A FIRST COURSE IN

STATISTICAL METH

BY

G. IRVING GAVETT

Assistant Professor of Mathematics, University of Washington

14

9-NO

COIM

FIRST EDITION

19

TUK

36

McGRAW-HILL BOOK COMPANY, INC.

NEW YORK: 370 SEVENTH AVENUE

LONDON: 6 & 8 BOUVERIE ST. E. C. 4

1925

McGraw-Hill Book Co. Inc.

PUBLISHERS OF BOOKS FOR

Electrical World ▿ Engineering [redacted]-Record
Power ▿ Engineering and Mining Journal-Press
Chemical and Metallurgical Engineering
Electric Railway Journal ▿ Coal Age
American Machinist ▿ Ingenieria Internacional
Electrical Merchandising ▿ Bus Transportation
Journal of Electricity and Western Industry
Industrial Engineer

Copyright, 1925, by the
McGraw-Hill Book Company, Inc.

PRINTED IN THE UNITED STATES OF AMERICA

1464

THE MAPLE PRESS COMPANY, YORK, PA.

PREFACE

This book is a growth from actual teaching experience in a course in Statistical Method with a large number of students in classroom and laboratory. This experience has spread over a period of about ten years. The content of the course has gradually been changed and the methods of presentation improved.

It is the aim of the book to serve as a text for a foundation course in Statistical Method for the different departments interested. The departments are mainly Mathematics, Business Administration, Sociology, Psychology, Commercial Engineering, Fisheries, and the Natural Sciences.

The presentation of the subject, as evolved, is that which the author has found to be workable. Only those things are presented which are deemed to be fundamental. After mastery of this text the student is well prepared to go on with specialized courses in any of the departments mentioned. He is also prepared to take up the more purely mathematical pursuit of the theory of probability and its applications to statistics.

No advanced mathematics is necessary for a satisfactory reading of the text. A very elementary idea of derivative is of assistance in the treatment of least squares adjustment and its application in deriving Pearson's coefficient of correlation. For the student unfamiliar with the idea of derivative, a simple explanation, satisfactory for the purpose desired, is given in Appendix D. However, by taking the meaning and results of a simple derivative on say-so, the application is readily made in the text.

In several cases, simplified derivations of formulas for statistical constants are presented. So far as the author has been able to find out, he has given the only absolutely correct definitions of median, quartiles, and other division points in print, that provide a logical derivation of a simple formula for determining them.

The exercises and illustrative examples are taken almost entirely from actual data. The exercises at the end of each chapter are merely suggestive. It has been found that the most satisfactory way of getting problems is to devise them as the

work proceeds. In this way the content of the problems is easily made to vary from term to term. The large newspaper almanacs, "The United States Statistical Abstract," and the "Commerce Yearbook" furnish an abundance of statistical material.

The author has used height and weight of freshman students at the University of Washington to illustrate the method of making frequency tables and their use in finding statistical constants. In order that each student may have his own material with which to work, lengths and breadths of a properly selected sample of leaves from a tree are found to give the most readily obtainable and satisfactory data.

The importance of the application of the principles of statistical method to most problems of human concern, where conclusions are to be drawn from a mass of observed data, is constantly being more thoroughly recognized. The necessity for the business man, as well as the biologist, the sociologist, and the psychologist, to be acquainted with the fundamentals of statistical method is becoming greater every day. Without a proper understanding of the underlying ideas involved, erroneous and sometimes dangerous results are employed.

The author's purpose in preparing this book has been to provide for the needs of his own students as shown by experience. If, in addition, the work may prove helpful to other instructors in giving a satisfactory foundation in the methods of statistics, he will be doubly repaid for his effort.

The author owes much to his coworkers for suggestions as to form and to his wife for putting the manuscript in shape for the publishers.

G. I. G.

SEATTLE, WASH.
November, 1925.

CONTENTS

		Page
Preface		v
Chap.		
I.	Introduction	1
II.	Tabulation, Frequency Distribution	19
III.	Graphical Representation. Frequency Graphs	42
IV.	Averages	96
V.	Dispersion	131
VI.	Skewness	169
VII.	Probable Error	177
VIII.	Curve Fitting	185
IX.	Correlation, Regression	212
X.	Logarithmic Graphical Representation	252
XI.	Index Numbers	273
Appendix A.	Logarithms	295
Appendix B.	Permutations, Combinations, Binomial Expansion	307
Appendix C.	Laws of Probability	315
Appendix D.	Derivatives and Integrals	321
Index		353

14

A FIRST COURSE IN STATISTICAL METHOD

CHAPTER I

INTRODUCTION

The noun *statistics* is either a singular noun or a plural noun according to the way in which it is used.

1. Definitions.—The "Standard Dictionary" gives the following definitions:

1. (Noun plural.) Numerical facts, collectively, pertaining to a body of things, especially where systematically gathered by direct enumeration and collated; specifically, such facts relating to a numerous body of people, as of a nation, state or social organization; as statistics of population; statistics of agriculture; church statistics; statistics for a census report.
2. (Noun singular.) The science that deals with the collection, classification, and tabulation, of such facts, especially as a branch of sociology.

The definitions given in Webster's "New International Dictionary" are as follows:

1. (Construed as singular noun.) The science of the collection and classification of facts on the basis of relative number or occurrence as a ground for induction; systematic compilation of instances for the inference of general truths.
2. (Construed as plural noun.) Classified facts respecting the condition in various respects of the people in a state, or respecting any particular class or interest; especially those facts which can be stated in numbers, or tables of numbers.

2. Discussion.—The definition of the plural noun as given by Webster is too narrow, as it confines itself to facts concerning the condition of people. That in the Standard includes facts pertaining to things in general. On the other hand, Webster's definition of the singular noun seems more explicit. Statistics is not merely the science of collecting, classifying, and tabulating facts. Its methods when applied to a set of facts properly selected from a

large body of facts enable us to draw conclusions by induction and establish general laws in regard to the large body of facts. The methods of statistics are employed by and are of great value to the physicist, the chemist, the biologist, the anthropologist, the meteorologist, the psychologist, the sociologist, the economist, the business man, and many others. The physicist and the chemist can frequently, by the experimental methods of the laboratory, segregate a given cause and observe directly its effect. The sociologist and the economist, on the other hand, can seldom isolate a given cause and determine directly the effect of that cause alone. He must make his observations on things as they exist in experience and not as they would be in a laboratory experiment. The methods of statistics are especially applicable to this situation. Even the observations of the physicist and the chemist are subject to human error and to variable causes not completely under control in the laboratory, and statistical methods become valuable again.

3. Yule's Definitions.—With these remarks in mind look at definitions given by Yule in his "Introduction to the Theory of Statistics." His work is one of the standard treatments of statistics in the English language. On page 5 he says:

By *statistics* we mean quantitative data affected to a marked extent by a multiplicity of causes.

By *statistical methods* we mean methods specially adapted to the elucidation of quantitative data affected by a multiplicity of causes.

By *theory of statistics* we mean the exposition of statistical methods.

4. Multiplicity of Causes.—These definitions indicate that statistical methods are especially applicable in those cases where an observed effect is due to a "multiplicity of causes."

An example is the yield of wheat per acre as affected by amount of rainfall. The yield of wheat per acre is due to a multiplicity of causes besides amount of rainfall. Among them are temperature, attacks by wheat rust, amount of fertilizer used, etc. The amount of fertilizer used may depend on its price and on the price of wheat. The price of wheat may depend on foreign demand. Foreign demand depends on something else. Statistical methods make it possible to study the effect of amount of rainfall as segregated from the other causes.

Another example is the effect of gravitation on the speed of a falling body. Friction of the air and a multiplicity of other things

besides gravitation affect the speed of a falling body. Statistical methods may be used. However, in this case the physicist may control the problem in his laboratory and reduce the effects of causes other than gravitation to nil or to negligible amounts. He can then observe directly the effect of gravitation.

5. Comparison.—One of the great purposes in a statistical investigation is *comparison.* One may wish to compare growth of population in one country with growth of population in another country, distribution of income in a certain country at one date with distribution of income at another date, heights of fathers and heights of their sons. These and many others are statistical problems.

6. Variables.—In each case the problem deals with variables. A *variable* is a quantity which changes in magnitude. It may change from one time to another, from one individual to another, or from one place to another. Distribution of income in the United States varies from one time to another. A variable which varies with time is called a *historical variable.* Height of fathers varies from one individual to another. Population of countries varies from one place to another.

Heights of men may vary through all possible numbers between certain limits. Such a variable is said to be a *continuous variable.* The number of spots coming up in throws of three dice vary from 3 to 18, but only by integral values. Such a variable is said to be a *discrete variable.* The observed values of the variable are called *variates, observations, size of item,* etc.

7. Average.—In order to make definite comparisons, it is necessary to get numerical values of certain constants which are functions of the particular variable in question. As an illustration take tree height as the variable. The heights of trees in one forest may be compared with the heights of trees in another forest by saying that the trees in the first forest are taller than those in the second. This does not necessarily mean that every tree in the first forest is taller than the tallest tree in the second. It is understood to mean that the trees of the first forest average taller than those of the second. A constant quantity called *average* is determined as representing tree height for the first forest. The same sort of average is determined for tree height of the second forest. It is difficult for the mind to grasp, at one time, the large number of tree heights in either forest. The two averages are readily compared. The average is a type form used to represent the height of the group as a whole.

8. Dispersion.—In comparing the tree heights in the two forests, it may happen that in the first forest, having the greater average tree height, there will be found very short trees and very tall trees, differing greatly from the average. At the same time in the other forest, the trees may be found to be nearly all of the same height, differing but little from the average. This might be of importance in estimating the amount of timber per acre for lumber. This scattering of values of the variable on each side of the average is called *dispersion*, or *variability*. A numerical measure of dispersion must be devised in order to more completely compare the tree heights of the two forests.

9. Tabulation and Classification.—If the height of each tree, taken in order, is measured and recorded, the resulting records of the two forests will be difficult to compare. This leads at once to tabulation and classification of the values of the variable into series, ordered according to size. When a set of observed values of a variable (variates) is arranged according to size in regular order from smallest to largest, or vice versa, the set or group is said to be *arraged*. The group then forms an *array*.

10. Sampling.—If there is a very large number of trees in the forest, it may become impossible or at least impracticable to measure every tree height. Then some method must be devised for selecting a tree here and a tree there, such that the average and the dispersion of tree heights of the group selected may be expected to be the same or nearly the same as for all the trees of the forest. This process is called *sampling*. The selected set of values of the variable is called a *sample*.

11. Statistical Regularity.—It is by virtue of what is known as *statistical regularity* that the study of a random sample of the entire group determines the properties of the entire group. Suppose it is desired to find the average height of all 7,000 students in a given university. Because of this law of statistical regularity, it is necessary to measure only a random sample of say about 700 heights. If the students all pass through a certain door and every tenth one is measured, it should accomplish the purpose. In selecting every tenth student, it should make no appreciable difference whether one starts with the first, the second, the fifth, or any one of the first ten. It should also give a fair sample if an alphabetical list of students is used and every tenth one in the list is chosen and called in to be measured. Again one might begin with the first in the list, or the second, or any one of the first ten.

For the sample to be truly *random*, the names of the 7,000 students may be written each on a card, the cards placed in a sack and thoroughly shaken up. Then a card may be drawn at random by a blindfolded person. The remaining cards in the sack are to be thoroughly shaken up again, and another one drawn at random. If the process is repeated until the required number of names for a good sample be thus selected, the heights of the students thus selected should constitute a random sample and be representative of the heights of the entire student body.

Various other ways of selecting a random sample may be devised. The necessary prerequisite is that the sample shall be representative of the entire group. The probability of large error diminishes with increase of the number of items in the sample. Use of a number of samples also diminishes the probability of large error. The mathematical theory of sampling cannot be taken up here.[1]

12. Classes.—Instead of tabulating and classifying all the tree heights of the forest, only those of the sample need to be used. When tabulating the values of the variable in the sample, they are usually divided into groups called *classes*. Each class covers the same range of values of the variable. This range is called the *class interval.* The number of variates in each class is called the *class frequency.* Such a table when completed is called a *frequency table.*

13. Graphs.—Graphical presentation of statistical facts is used very extensively. The construction and interpretation of graphs or charts is a complete study by itself. Many of the problems of the statistician can hardly be solved without graphs; most of them are much clarified by the use of graphs. By using graphs many comparisons are easily made by the eye. A study of economic situations and business conditions is much facilitated by use of graphs. The graphs of frequency tables are called *frequency graphs* or *histograms.* The graph of a historical variable is called a *historigram.* Do not confuse this term with histogram.

14. Deviations, Symmetry.—In studying dispersion, the amount each variate differs from the average is found. These differences are called *deviations* from the average. If the variate is less than the average, the deviation is negative; if greater than the average, the deviation is positive. If, for each positive

[1] See G. Udney Yule, "Introduction to the Theory of Statistics," part III.

deviation, there is a negative deviation of the same size, the distribution with respect to the average is *symmetrical.* The occurrence of asymmetry leads to the study of *skewness.* In the first case the frequency graph is symmetrical, usually bell-shaped. If skewness is present, the frequency graph is bell-shaped with one extremity pulled out along the base line farther than the other. Besides these two type forms of frequency curves, there are a number of others which can be thoroughly understood only by a study of *curve-fitting.*

15. Correlation.—If a comparison be made of two variables such as length and breadth of a given variety of tree leaves, or amount of employment and savings-bank deposits, it would be expected that with increase of leaf length there will, in general, be an increase of leaf breadth, or that with change of amount of employment there will be a corresponding change, in the same direction, in savings-bank deposits. If the supply of a commodity increases, it would be expected that the price would go down. When corresponding values of two variables seem to have some connection so that when one increases, the other has a tendency to increase, or so that when one increases, the other has a tendency to decrease, there is said to be *correlation* between the two variables. The numerical measure of the extent to which certain values of one variable tend to be associated with a selected value of the other is called *coefficient of correlation.*

16. Use of Formulas.—The above considerations will be followed to a large extent in the succeeding chapters. Some modifications may be employed and some topics be introduced that have not been mentioned. Formulas will be derived for finding the numerical values of these various statistical constants. Many formulas have limitations in their application. Sometimes the method of derivation of a formula makes its limitations obvious. Knowledge of the application of a formula cannot be considered complete unless its limitations are known. Do not follow a formula blindly. Know what it means, what it signifies, what knowledge it is intended to convey. If possible have a thorough understanding of its derivation. Otherwise incorrect conclusions may be drawn. In such case results may be obviously wrong, or they may not be detected until irreparable damage is done. Thus distrust of statistics and statistical method is aroused.

17. Accuracy, Errors.—Other sources of distrust are inaccurate data and inaccurate computations. Absolutely exact measure-

ments of physical things are unattainable. If a surveyor measures the length of a base line with the utmost care, he gets a number of different results. This is due to the imperfections of the apparatus with which the measuring is done and to the *inability of man to apply the method always in exactly the same* way. The exact value of the measurement is unknown. What is known as the most probable value is used. If there is no reason for thinking one measure is better than another, the arithmetic mean is used as the most probable value. The actual measures, or observations, will be some of them larger and some of them smaller than the arithmetic mean. The difference between an observed value and the arithmetic mean is called an *error* or *deviation from the mean.* These errors, being some positive and some negative, tend to offset each other, and are called *compensating errors.* On the other hand, if the line had been measured with a hundred-foot tape which was too short, then the longer the line measured the greater the error due to defective tape length. Such errors are called *cumulative errors.* Compensating errors can be largely taken care of by having a great number of measurements. Cumulative errors become worse the oftener it is necessary to apply the measuring stick.

18. Some so-called statisticians have been known to say that inaccuracy in the data makes no difference if there is only a sufficiently large amount of data. The old man who bought eggs at 25 cts. a dozen and sold them at 24 was asked how he could make any money. He replied that he could not, except that he sold a very great many. If it be true that when a woman is asked her age she is inclined to understate it, the correct average age of women cannot be obtained by asking even a very large number of women how old they are. But if it be true that some *women are inclined to understate and others to overstate their* ages, it may happen that one can arrive very closely at the correct average by getting a sufficiently large number of women to state their ages.

19. Mistakes.—Beside the sorts of errors just discussed, there are what may be called mistakes. These should not be allowed to pass. A constant checking of work in every way possible should eliminate mistakes. It is not a crime to make a mistake. Everybody does it, but it is, at least, very reprehensible to let the mistake stand. A mistake in an engineer's computations sometimes results in disaster with great loss of life.

20. Standard or Degree of Accuracy.—In every problem or investigation involving numerical quantities, there should be a standard or degree of accuracy established. Every item should be brought to this standard. This does not mean that every measurement must be the most accurate possible. It may serve the purpose of a statistical problem to know the population of a state to the nearest ten thousand or it may be needed to the nearest hundred. The number of students in the various classes of the university might be needed to the nearest unit. The standard of accuracy needed is determined by the purpose for which the data are to be used. The standard of accuracy is limited by the closeness with which the quantities involved can be measured.

21. Recording of Numbers.—Numbers are recorded so that each shows the standard of accuracy used. If populations of cities are used to the nearest thousand, then each number is recorded with three ciphers at the end to fill out to the decimal point. Thus:

City	Population
A	427,000
B	536,000
C	720,000
D	329,000

Sometimes these ciphers that serve merely as fillers are written smaller than the other figures. In the above table the use of fillers could be avoided by indicating the standard of accuracy in the column heading.

City	Population, in Thousands
A	427
B	536
C	720
D	329

If the lengths of table tops are being measured to the nearest hundredth of a foot, then record each result to show that degree of accuracy. If one table top seems to measure 5 ft., do not record it as 5, but write the result 5.00, using ciphers to fill in to hundredths place.

22. False Accuracy.—Certain numerical processes may be carried out to indicate a higher degree of accuracy than that actually attained. This should not be done. Usually, the result is no more accurate than the least accurate element in the computation. This may be illustrated by an example in multiplica-

tion. Suppose the length and the breadth of a table top, measured to the nearest hundredth of a foot, are 5.21 and 2.34 ft., respectively. Find the area in square feet. The data have three-figure accuracy. No greater than three-figure accuracy can be expected in the result. Multiplying in the ordinary way gives:

```
   5.21
   2.34
  -----
   2084
  1563*
 1042**
 ------
12.1914
```

Using three-figure accuracy this result would be read 12.2 sq. ft.

The length given as 5.21 ft. means that the actual length was somewhere between 5.205 and 5.215 ft. In the same way the actual breadth is somewhere between 2.335 and 2.345 ft. The product of the smaller length and breadth, 5.205 and 2.335, is 12.153675. The product of the greater length and breadth, 5.215 and 2.345, is 12.229175. The area then lies between these two products. They agree if the right-hand figures are dropped till only three-figure accuracy is indicated. The result is 12.2, which agrees to three figures with the product of the original dimensions as measured.

23. By inspecting the original multiplication of 5.21 by 2.34, it is readily seen that the two right-hand figures, 14, are very likely incorrect and that the 9 is somewhat unreliable. It is not known whether the 5.21 is too large or too small. So the last figure, 4, of the first partial product is uncertain. The same thing is true of the last figures of the other partial products. Moreover, if the measurement of length and breadth had been carried out to the nearest thousandth of a foot, one would not have the slightest idea what figures would replace the *'s shown in the example. It is thus very evident that the last two figures, 14, of the product are almost sure to be wrong and that the 9 is at least somewhat uncertain. Writing the result as 12.1914 sq. ft. tells a lie. It leads one to believe that there is six-figure accuracy in the result, where there really is only three-figure accuracy. Similar considerations apply to division, square root, and other processes of computation.

24. Aids to computation should be freely used. Crelle's multiplication tables, Barlow's tables of squares, square roots, etc., and tables of logarithms are very useful. The common

pocket slide rule gives very quick results where only three-figure accuracy is necessary. This is usually the case in computations of percentages. If any extended statistical work is to be done, a computing machine is almost indispensable.

TABLE I.—WEIGHT AND HEIGHT OF FRESHMAN MEN AT THE UNIVERSITY OF WASHINGTON, 1923–1924

Weight in pounds. Height in inches

Weight	Height	Weight	Height	Weight	Height	Weight	Height	Weight	Height	Weight	Height
157	75.0	167	70.0	162	68.0	137	66.0	150	70.0	170	70.0
134	67.0	147	68.5	168	68.5	102	63.5	153	68.0	135	65.0
133	65.5	130	65.5	125	66.0	138	68.0	126	69.0	136	64.5
125	71.0	153	67.5	155	72.0	155	70 5	143	71.0	137	63.5
123	68.0	160	67.5	120	63.5	133	65.5	139	72.0	148	69.0
136	68.0	147	72.0	156	68.0	143	67.5	169	71.5	124	63.0
163	70.0	157	69.5	116	63.5	134	68.0	135	70.0	117	64.5
114	70 0	165	69.0	108	64.0	155	72.0	119	66.0	128	70.0
135	68.5	122	65.0	151	68.0	141	66.0	190	69.0	143	70.5
157	71.5	142	67.5	178	72.5	180	72.0	141	70.0	129	62.0
145	68.0	131	69.5	137	67.5	137	68.5	158	70.0	163	70 0
151	72.0	141	68.5	125	65.5	147	71.0	146	70.0	139	67.0
174	67.5	136	68.5	130	69.5	155	71.0	122	68.5	122	62.5
131	69.0	152	68.0	162	70.0	178	06.0	176	74.5	134	67.5
161	72.0	151	69.5	123	66.5	169	73.0	139	70 5	140	67.0
149	68.0	143	68.0	125	67.0	134	69.5	126	65.0	132	68.5
151	70.0	159	71.0	122	68.0	131	70.5	133	68.0	120	65.5
144	67.5	163	75.0	151	08.5	164	72.5	133	71.0	148	68.0
148	67.0	145	67.0	139	67.0	149	67.0	160	67.5	129	67.0
133	66.5	147	68.5	132	66.0	155	69.5	132	07.0	152	67.0
118	70.0	136	69.0	140	70.0	169	66.5	163	71.0	150	64.0
136	65.0	138	65.5	165	69.5	137	67.5	121	63.5	125	64.0
142	69.5	148	67.0	124	66.0	162	70.5	110	67.0	178	72.0
163	70.5	160	71.0	126	65.0	130	67.5	130	64.5	112	67.0
145	67.5	162	70.0	133	69.0	132	70.5	148	69.0	130	68.5
141	70.0	143	69.0	144	69.5	122	66.5	136	66.5	128	68.0
130	68.5	150	72.0	127	66.0	137	71.5	116	67.0	115	65.5
138	68.5	130	66.5	158	72.0	154	72.0	109	70.5	144	65.5
150	67.0	141	68.0	142	70.0	141	69.0	136	71.0	173	72.0
149	68.5	141	67.5	149	69.0	137	72.0	161	70.0	138	68.0
155	66.0	138	68.0	145	68.5	155	69.5	138	69.0	130	69.0
153	69.5	133	66.5	150	68.0	120	66.0	125	64.5	138	67.0
124	68.0	141	72.0	120	64.5	135	70.0	101	62.5	143	70.5
129	71.0	112	60.5	145	67.5	128	69.0	124	69.0	120	63.5
173	69.5	131	67.0	133	63.0	140	69.0	146	67.6	132	69.0
155	69.5	151	69.0	139	67.5	141	67.0	139	71.0	141	67.5
165	72.0	163	70.5	132	68.0	139	68.0	127	65.5	129	66.0
142	67.0	124	65.5	178	66.0	130	64.5	183	71.0	133	66.5
120	67.0	126	64.0	174	72.5	152	67.0	131	68.5	163	73.0
134	71.0	130	63.5	181	69.0	153	72.0	141	67.5	157	67.5
150	69.0	172	69.0	119	65.5	164	69.0	174	70.5	176	72.0
125	67.0	131	65.0	175	73.0	146	66.5	134	68.0	133	68.0
165	67.5	170	64.0	133	66.0	138	65.5	132	67.0	136	62.0
141	70.5	135	64.5	171	73.0	128	70.0	122	63.5	135	67.5
125	70.0	132	66.0	133	68.0	142	68.0	191	72.5	145	67.5
142	67.5	140	69.0	138	69.0	161	67.0	110	64.0	124	66.5
149	64.0	132	68.0	140	69.5	145	68.5	140	67.0	120	69.0
160	67.5	130	65.0	122	66.0	135	69.5	124	68.0	145	69.0
141	70.0	160	69.0	138	72.0	117	65.5	129	65.0	180	70.5
145	67.5	153	69.0	178	70.5	125	64.0	140	69.0	184	68.0

TABLE I.—(*Continued*)

Weight	Height	Weight	Height	Weight	Height	Weight	Height	Weight	Height	Weight	Height
178	69.5	127	66.5	126	67.5	145	69.5	135	70.0	150	68.5
144	67.0	156	70.5	137	67.0	121	68.0	159	70.0	145	68.5
145	67.5	120	66.0	150	67.5	167	67.0	165	67.0	154	69.5
146	69.0	145	69.0	128	67.5	136	70.0	144	68.5	171	72.0
159	73.0	131	66.0	138	68.0	125	68.5	120	66.5	136	67.5
156	70.5	142	69.0	125	63.5	114	64.5	129	68.5	107	68.5
148	69.0	144	70.0	117	65.5	149	67.5	142	72.0	143	68.5
151	69.5	205	69.5	140	70.0	127	69.5	132	63.5	137	67.5
131	69.5	136	67.5	143	70.0	159	66.0	130	70.0	152	70.0
153	69.5	142	66.0	129	63.0	112	63.5	161	71.0	160	67.5
150	70.0	142	69.0	116	65.5	132	69.0	141	67.0	172	70.5
167	74.5	130	68.5	145	71.0	146	70.0	115	66.5	140	68.5
152	69.0	117	67.5	162	73.0	116	64.0	157	67.5	131	68.0
119	67.5	145	60.0	134	67.0	140	66.0	124	66.0	154	68.0
148	60.0	137	66.5	132	67.0	146	65.0	156	69.0	116	69.0
120	65.5	140	69.0	120	65.0	152	68.5	208	70.0	123	66.5
136	70.0	123	67.5	143	70.0	144	68.5	132	68.0	145	69.0
144	72.0	124	64.0	145	69.0	128	70.0	216	69.0	139	68.5
132	67.5	199	68.0	162	66.5	170	68.5	126	68.0	147	66.5
131	66.5	179	71.0	137	66.5	115	64.5	134	68.0	153	70.0
194	68.5	149	69.5	122	64.0	130	67.0	136	65.0	133	67.0
111	65.0	129	65.0	156	68.0	135	70.0	110	67.5	174	71.0
122	68.5	140	72.5	194	70.5	150	67.0	143	67.5	155	67.5
124	63.0	119	70.0	172	75.0	147	72.0	166	68.5	136	68.5
145	72.5	144	70.5	155	71.0	142	68.5	148	67.5	158	72.0
125	71.0	124	69.5	155	71.0	112	67.0	162	71.0	122	69.0
122	68.0	117	65.0	115	69.0	152	70.0	157	70.0	158	74.5
133	65.5	137	66.0	174	69.5	127	66.0	131	69.0	130	65.0
159	67.0	162	71.5	148	68.0	143	66.0	139	65.5	149	70.0
100	66.5	154	71.5	119	65.5	147	68.0	142	68.5	130	70.0
125	71.5	109	66.0	156	71.0	130	66.0	142	68.0	135	61.0
126	67.5	148	68.0	150	68.0	155	71.0	136	66.5	153	70.5
168	73.5	129	66.0	163	70.0	145	67.5	145	66.5	145	67.0
149	70.0	157	69.0	153	67.0	110	67.0	185	66.5	141	67.0
142	68.0	141	65.5	122	67.0	133	65.0	126	70.5	142	68.5
167	70.0	175	69.0	159	71.0	161	72.5	135	65.5	122	66.5
153	71.0	115	67.5	153	70.0	125	66.0	132	66.0	179	69.5
152	67.5	144	64.0	141	69.0	107	67.5	187	70.5	135	69.5
120	60.5	136	67.5	123	62.0	112	64.0	152	70.5	126	67.5
137	72.5	138	67.0	146	64.0	182	66.5	136	62.5	143	69.5
161	69.0	150	64.0	144	70.5	140	67.5	122	64.0	137	67.5
136	68.0	131	69.0	122	67.0	145	70.5	163	70.5	126	65.5
111	61.0	139	67.0	157	68.5	118	67.5	148	67.0	125	65.0
139	66.5	173	74.0	110	62.0	161	69.0	149	70.5	151	69.5
130	68.5	148	65.0	120	66.5	161	72.0	126	67.5	140	69.0
139	68.5	143	69.0	169	68.0	145	72.5	162	70.0	130	70.5
149	67.0	165	71.0	153	68.0	109	62.0	163	69.0	144	69.5
170	72.0	194	69.5	136	66.5	156	67.0	139	70.0	170	71.5
145	66.5	152	68.0	141	66.0	130	67.0	158	68.0	142	67.5
153	67.5	165	69.0	131	69.0	129	67.0	123	68.5	136	66.0
138	64.0	142	69.0	153	66.0	125	66.0	150	73.0	126	63.5
126	69.5	105	67.5	155	68.0	148	67.5	146	70.5	138	69.5
135	68.5	147	69.0	129	68.0	130	68.5	145	71.0	153	70.5
136	66.0	140	70.0	153	68.0	147	69.0	126	65.0	125	69.5
141	69.0	176	68.5	125	65.0	172	72.0	163	72.5		

TABLE II.—WEIGHTS AND HEIGHTS OF 629 FRESHMAN MEN AT THE UNIVERSITY OF WASHINGTON

Weight, in pounds	Height, in inches																														Totals
	60.5	61.0	61.5	62.0	62.5	63.0	63.5	64.0	64.5	65.0	65.5	66.0	66.5	67.0	67.5	68.0	68.5	69.0	69.5	70.0	70.5	71.0	71.5	72.0	72.5	73.0	73.5	74.0	74.5	75.0	
100													1																		1
101					1																										1
102							1																								1
103																															0
104																															0
105															1																1
106																															0
107															1		1														2
108								1																							1
109				1								1									1										3
110				1				1						2	1																5
111		1								1																					2
112	1						1	1						2																	5
113																															0
114									1											1											2
115									1		1		1		1			1													5
116							1	1			1			1				1													5
117									1	1	2				1																5
118															1					1											2
119											2	1			1					1											5
120							1		1	1	2	2	2					1													10
121							1									1															2
122					1		1	2		1		1	2	2		2	2	1													15
123				1									2		1	1	1														6
124						2		1			1	2	1			2		1	1												11
125							1	2	1	2	1	3		2			1		1	1		2	1								18
126							1	1		3	1		1		4	1		1	1		1										15
127											1	2	1						1												5
128															1	1		1		3											6
129				1		1	1			2		2		3		1	1					1									13
130							1		2	2	1	1	1	2	1		5	1	1	2	1										21
131										1		1	1	1		1	1	4	2		1										13
132							1					3		3	1	3	1	2			1										15
133						1				1	3	1	3	1		3		1				1									15
134														2	1	3			1			1									8
135		1							1	1	1				1		2		2	4											13
136				1	1				1	2		2	3		3	2	2	1		2		1									21
137							1					2	2	1	4		1						1	1	1						14
138								1			2			2		4	1	2	1					1							14
139											1		1	3	1	1	2			1	1	1		1							13
140												1		2	1		1	5	1	3					1						15
141											1	2		3	3	1	1	3		3	1			1							19
142												1		1	3	3	3	3	1	1				1							17
143												1			2	1	1	2	1	2	2	1									13
144								1			1			1	1		2		2	1	2			1							12
145													2	2	6	1	3	5	1		1	2			2						25
146								1		1			1		1			1		2	1										8
147													1			1	2	2				1		2							9
148										1				3	2	3		4													13
149								1						2	1	1	1	1	1	2	1										11
150								2						2	1	2	1	1		2				1		1					13
151																1	1	1	3	1				1							8
152														2	1	2	1	1		2	1										10
153												1		1	2	3		1	2	2	2	1		1							16
154																1			1				1	1							4
155												1			1	1			3		1	4		2							13
156														1		2		1			2	1									7
157															2		1	1	1	1			1							1	8
158																1				1				2					1		5

TABLE II.—(*Continued*)

Weight, *in* pounds	Height, in inches																														Totals
	60.5	61.0	61.5	62.0	62.5	63.0	63.5	64.0	64.5	65.0	65.5	66.0	66.5	67.0	67.5	68.0	68.5	69.0	69.5	70.0	70.5	71.0	71.5	72.0	72.5	73.0	73.5	74.0	74.5	75.0	
159												1		1						1		2				1					6
160															4			1				1									6
161														1				2		1		1		2	1						8
162													*1*			*1*				*3*	*1*	*1*	*1*			*1*					*9*
163																		1		3	3	1			1	1				1	11
164																		1							1						2
165														1	1			2	1			1		1							7
166																	1														1
167														*1*						*2*									*1*		*4*
168																	1										1				2
169													1			1							1			1					4
170								1									1			1			1	1							5
171																								1		1				1	3
172																		*1*			*1*			*1*							*3*
173																			1					1				1			3
174															1				1		1	1			1						5
175																		1								1					2
176																	1							1					1		3
177																															*0*
178												2							1		1			1	1						6
179																			1			1									2
180																					1			1							2
181																		1													1
182													*1*																		*1*
183																						1									1
184																1															1
185													1																		1
186																															0
187																					*1*										*1*
188																															0
189																															0
190																		1													1
191																									1						1
192																															*0*
193																															*0*
194																	1		1		1										3
195																															0
196																															0
197																															*0*
198																															0
199																1															1
200																															0
201																															0
202																															*0*
203																															0
204																															0
205																			1												1
206																															0
207																															*0*
208																				1											1
209																															0
210																															0
211																															0
212																															*0*
213																															0
214																															0
215																															0
216																		1													1
Totals..	1	2	0	5	3	4	12	17	9	20	22	34	30	51	58	54	44	62	35	51	30	27	7	26	10	7	1	1	2	2	629

Table III.—Monthly and Annual Precipitation[1] at Seattle

Year	January	February	March	April	May	June	July	August	September	October	November	December	Annual
1900	3.04	4.35	4 45	1.55	**3.73**	2.51	0.66	0.30	0.72	4.16	3.80	7.21	36.48
1901	4 26	4.26	1.62	3 86	1.44	1.90	0.35	0.13	2.30	1.44	6.17	2 45	30.18
1902	5.25	**8.10**	4.19	2.20	1.86	1.71	**2.01**	0.48	2.15	2.71	6.30	8.82	**45.76**
1903	3.12	1.45	6.10	1.56	2.39	1.55	1.28	0 50	3.18	1.94	7.34	4.14	34.55
1904	4.26	5 31	**6.22**	3.08	**0.54**	1.23	0.58	0.07	0.22	2.32	**8.61**	5.29	37.73
1905	5.61	2.76	3.45	1.04	3.37	3.03	0.36	0.44	1.70	4.29	2.68	5.62	34.35
1906	5.03	4.60	0.89	**0.34**	2.64	1.97	0.11	0.15	3.31	3.16	7.67	6.80	36 67
1907	4.18	3.87	1 12	1.54	0.98	0.78	0.71	0 80	**3.39**	0.67	4.73	6.33	29.10
1908	4.10	4.24	2.54	2.10	3.19	0.15	0.24	0.82	0.23	2.34	4.60	3.61	28.25
1909	6 90	4.35	1 08	0.77	1 60	0.64	0 35	0.27	1.10	2.86	9 11	2.69	31.72
1910	5 08	5.03	1.80	2.41	1.88	0.82	0 01	0.17	1.04	4.02	8.47	3.47	34.20
1911	3 67	1.42	**0.88**	1.21	2.48	0.44	0.18	0.13	3.27	1.00	3.26	3.75	**21.69**
1912	4 52	3 11	1.79	1.73	1.64	2.76	1.15	**2.49**	0.73	3.97	6 82	4.43	35.14
1913	4 89	1.34	1.55	0.83	1 37	1.71	0.73	0 45	2.37	2.00	4.74	2.61	24 59
1914	**9.82**	1.93	1.40	3 31	0.74	1.75	0.01	**0.01**	1.42	**4.37**	5.28	**1.39**	31.43
1915	6 35	2.76	1.72	2.91	1.72	0.40	0.84	0.05	0.65	3.00	5.66	7.77	33 83
1916	4.32	6.85	5.45	1.98	1.56	1.82	1.93	0.11	0 70	1.18	4.58	4.13	34.61
1917	2.02	1 43	2.96	**4.48**	0.83	**3.70**	0.09	0.03	1.29	**0.16**	2.70	**9.21**	28.90
1918	2.94	4.81	3.92	0.96	1.19	0.50	1.38	1.12	**0.08**	3.46	3.81	5.04	29.21
1919	7.95	3.77	1.84	3.20	2.08	0.35	0.22	0.08	2.03	1.50	4.13	4 10	31.34
1920	3.92	**0.34**	2.82	3.46	0.96	1.93	1.00	1.15	2.34	4.19	4.42	5.68	32.21
1921	5.56	4.82	3.06	1.76	1.93	1.29	0.18	1.61	1.84	3.91	6.60	7.25	39.81
1922	**1.89**	1.74	4.45	2.53	1.08	**0.03**	**0.00**	1.17	1.19	2.37	**1.45**	7.37	25.27
1923	7.51	2.72	1.37	1.67	1.45	1.01	0.68	1.98	1.37	2.05	2.06	3.31	27.18
Average	4.84	3.56	2.77	2.11	1.77	1.42	0.63	0.60	1.61	2 63	5.22	5.10	32.26

Bold-faced figures indicate greatest and least monthly and annual amounts.

[1] Annual Meteorological Summary of the Weather Bureau Office at Seattle, Wash., M. B. Summers, Meteorologist.

TABLE IV.—MONTHLY AND ANNUAL MEAN TEMPERATURE[1] AT SEATTLE

Year	January	February	March	April	May	June	July	August	September	October	November	December	Annual
1900	43.8	42.3	49.9	51.8	56.0	61.9	65.0	62.6	58.6	51.8	44.4	45.8	52.8
1901	40.2	42.9	46.0	47.8	55.4	57.1	61.6	65.7	58.0	57.2	49.2	42.4	52.0
1902	39.8	46.4	45.0	50.0	56.5	60.6	63.8	64.4	58.4	54.2	45.8	41.8	52.2
1903	42.9	40.2	42.5	47.9	54.6	61.9	62.0	63.2	57.7	53.6	46.0	42.6	51.2
1904	41.8	40.2	42.6	52.6	55.1	59.7	64.2	63.6	61.0	54.2	49.4	43.0	52.4
1905	40.8	43.3	49.8	52.7	54.5	59.2	64.6	62.6	59.4	48.6	45.2	42.6	51.0
1906	42.4	43.8	45.5	53.0	54.6	58.0	67.7	65.0	58.9	54.0	45.2	42.4	52.5
1907	33.8	41.8	43.2	49.1	57.0	59.4	64.2	61.3	58.6	53.2	48.0	43.4	51.1
1908	41.8	41.7	44.2	49.0	51.6	58.8	65.1	62.2	57.1	50.8	48.8	39.5	50.9
1909	34.2	42.0	44.1	47.7	52.3	59.4	60.6	61.6	60.1	52.2	45.5	36.4	49.7
1910	39.0	38.3	47.6	49.4	57.0	57.5	62.6	60.2	58.0	52.8	45.5	42.8	50.9
1911	38.0	39.8	45.8	46.6	52.7	57.0	65.0	62.6	57.1	52.5	44.7	41.2	50.2
1912	42.6	43.9	44.3	48.0	57.0	60.0	63.4	62.2	59.2	49.0	46.2	41.6	51.5
1913	36.6	40.0	41.9	49.0	54.1	59.5	63.4	64.8	58.8	50.1	46.2	42.4	50.6
1914	43.2	42.3	47.6	51.4	57.3	58.9	64.2	63.2	56.7	54.6	47.2	39.7	52.2
1915	40.6	44.5	50.0	52.6	56.0	59.8	64.3	66.8	59.1	53.7	43.7	42.0	52.8
1916	31.0	41.9	44.4	49.0	52.0	58.8	61.1	63.6	58.8	49.1	43.0	38.0	49.2
1917	38.0	39.3	41.0	46.8	52.4	57.2	63.4	65.2	58.9	52.5	49.6	45.0	50.8
1918	43.7	40.0	44.0	50.0	52.4	61.6	63.0	62.6	62.2	53.4	45.8	40.9	51.6
1919	41.4	40.8	44.7	49.6	53.6	57.5	63.0	63.0	59.6	48.5	44.9	38.6	50.4
1920	40.2	40.3	44.4	45.6	51.6	58.6	64.2	64.4	57.8	50.1	47.0	43.4	50.6
1921	40.4	42.9	44.6	47.5	53.6	59.8	60.8	62.0	57.0	53.2	45.4	39.1	50.5
1922	35.5	39.4	41.5	46.6	54.5	60.8	62.9	62.7	59.8	53.5	43.6	38.4	49.9
1923	40.4	37.3	44.0	51.0	54.1	60.6	64.4	65.7	60.8	54.6	47.4	42.2	51.9
Average	39.7	41.4	44.9	49.4	54.4	59.3	63.5	63.4	58.8	52.4	46.2	41.5	51.3

[1] Annual Meteorological Summary at the Weather Bureau Office at Seattle, Wash., M. B. Summers, Meteorologist.

TABLE V.—NUMBER OF HEADS AND TAILS FALLING IN EACH OF 500 THROWS OF SEVEN DIMES

Heads	Tails	Heads	Tails	Heads	Tails	Heads	Tails	Heads	Tails	Heads	Tails	Heads	Tails	Heads	Tails	Heads	Tails	Heads	Tails
2	5	4	3	3	4	3	4	4	3	1	6	3	4	4	3	3	4	3	4
2	5	2	5	3	4	2	5	1	6	4	3	5	2	4	3	3	4	4	3
2	5	2	5	1	6	2	5	3	4	3	4	3	4	2	5	6	1	3	4
4	3	4	3	4	3	4	3	3	4	5	2	3	4	4	3	3	4	3	4
1	6	6	1	6	1	2	5	6	1	6	1	4	3	4	3	6	1	2	5
3	4	4	3	4	3	5	2	3	4	5	2	4	3	6	1	3	4		
5	2	2	5	4	3	3	4	3	4	5	2	5	2	3	4	3	4		
3	4	4	3	3	4	4	3	3	4	4	3	4	3	4	3	3	4		
4	3	1	3	4	3	3	4	2	5	4	3	3	4	5	2	3	4		
4	3	4	3	4	3	5	2	3	4	2	5	2	5	3	4	2	5		
5	2	2	5	4	3	5	2	3	4	1	6	2	5	1	6	3	4		
5	2	5	2	2	5	2	5	1	6	2	5	2	5	3	4	2	5		
1	6	3	4	6	1	3	4	4	3	2	5	4	3	3	4	2	5		
4	8	5	2	4	3	3	4	4	3	5	2	0	7	3	4	5	2		
2	5	5	2	4	3	2	5	1	6	3	4	5	2	3	4	4	3		
1	6	3	4	4	3	2	5	5	2	5	2	3	4	4	3	3	4		
3	4	2	5	4	3	4	3	4	3	4	3	5	2	2	5	2	5		
1	6	6	1	6	1	3	4	1	6	3	4	1	6	3	4	2	5		
3	4	2	5	4	3	3	4	2	5	2	5	4	3	4	3	3	4		
3	4	2	5	4	3	3	4	2	5	6	1	5	2	2	5	3	4		
3	4	1	6	4	3	3	4	4	3	4	3	1	6	3	4	2	5		
5	2	3	4	2	5	5	2	3	4	2	5	4	3	5	2	1	6		
3	4	4	3	2	5	3	4	4	3	4	3	2	5	6	1	3	4		
4	3	4	3	3	4	4	3	3	4	2	5	5	2	6	1	3	4		
4	3	3	4	4	3	3	4	5	2	2	5	1	6	4	3	4	3		
5	2	4	3	1	6	4	3	3	4	4	3	3	4	3	4	4	3		
4	3	2	5	3	4	3	4	3	4	5	2	3	4	5	2	4	3		
1	6	3	4	3	4	3	4	2	5	4	3	4	3	3	4	3	4		
5	2	3	4	4	3	2	5	4	3	3	4	4	3	4	3	4	3		
3	4	5	2	5	2	4	3	4	3	2	5	2	5	4	3	3	4		
5	2	3	4	3	4	3	4	3	4	5	2	3	4	3	4	4	3		
5	2	5	2	5	2	3	4	3	4	2	5	4	3	2	5	6	1		
5	2	5	2	5	2	4	3	4	3	4	3	3	4	3	4	5	2		
2	5	4	3	2	5	5	2	3	4	3	4	4	3	2	5	3	4		
4	3	4	3	4	3	1	6	5	2	6	1	4	3	5	2	1	6		
4	3	3	4	4	3	4	3	5	2	4	3	4	3	3	4	5	2		
5	2	2	5	5	2	2	5	3	4	4	3	4	3	7	0	3	4		
5	2	6	1	5	2	2	5	5	2	4	3	5	2	1	6	4	3		
3	4	4	3	3	4	3	4	5	2	3	4	5	2	6	1	3	4		
5	2	3	4	4	3	2	5	6	1	2	5	4	3	3	4	4	3		
4	3	3	4	5	2	3	4	5	2	4	3	3	4	2	5	3	4		
4	3	4	3	2	5	5	2	2	5	4	3	5	2	4	3	5	2		
3	4	5	2	4	3	2	5	4	3	4	3	1	6	4	3	3	4		
4	3	2	5	5	2	3	4	3	4	4	3	5	2	3	4	3	4		
1	6	3	4	1	6	4	3	2	5	4	3	4	3	3	4	4	3		
2	5	4	3	3	4	5	2	6	1	3	4	4	3	4	3	5	2		
1	6	4	3	3	4	2	5	3	4	4	3	3	4	4	3	5	2		
2	5	5	2	6	1	4	3	4	3	3	4	4	3	4	3	3	4		
3	4	4	3	3	4	3	4	3	4	4	3	1	6	3	4	4	3		
3	4	6	1	2	5	4	3	3	4	2	5	4	3	5	2	3	4		
3	4	5	2	3	4	5	2	3	4	3	4	3	4	1	6	5	2		
3	4	4	3	2	5	5	2	2	5	2	5	5	2	4	3	2	5		
4	3	4	3	6	1	3	4	1	6	2	5	4	3	4	3	4	3		
3	4	6	1	1	6	4	3	3	4	4	3	7	0	3	4	7	0		
3	4	1	6	3	4	5	2	4	3	2	5	4	3	5	2	4	3		

Exercises

1. In Table I are recorded the height and weight of 629 freshman men in the year 1923–1924 at the University of Washington. These records were selected from the University Health Service cards by taking every other card as arranged in alphabetic order.

From this table is it easy to determine the average height and average weight of freshmen? Is it easy to determine the most common height and the most common weight?

2. Construct Table II as follows:

(*a*) By examining Table I the largest and smallest values of height and weight are determined.

(*b*) A form like that of Table II is prepared by entering in the first column all possible values of the variable weight from smallest to largest and in a row across the top all possible values of the variable height from smallest to largest. Then horizontal and vertical lines are drawn between the values of the variates dividing the sheet of paper into cells, one cell for each value of weight and each value of height that may be associated with it.

(*c*) Beginning with the first entry in Table I, make a score mark in Table II *in the cell corresponding to this first entry of* Table I. *That is, make a score* mark in the cell that is in the column headed 75.0 and in the line headed 157. Proceed in like manner with each entry of Table I.

(*d*) Prepare a final form like that of Table II, entering in each cell the number of score marks in the corresponding cell of the table obtained in (*c*).

(*e*) Form the right-hand column of totals of entries in each line, and the bottom row of totals of entries in each column. See that the total of the column of totals and of the line of totals is the same and equal to the total number of pairs of entries in Table I. If this is not so, find the mistake and correct it.

3. From Table II is it any easier to find the average height and the average weight of freshmen? Why?

Is it easier to determine the most common height and the most common weight? Why?

4. Which is more common, small deviations from the average height or large deviations from the average height? Explain how determined. Is the same answer true in regard to weight?

5. Is there any evidence that tall men have a tendency to be heavy and short men light? What shows this?

6. If you were given the height of some freshman, could you, by means of this table, predict his weight either exactly or approximately? On what considerations do you base your answer?

7. As shown by Table II, about what weight would you expect a freshman of height 71.0 in. to be?

8. By inspection of Table III ascertain in which months Seattle has the greatest precipitation. In which months the least.

9. What is the average for each of the two months of greatest precipitation during the period 1900–1923 inclusive?

What is the average for each of the two months of least precipitation?

10. From Table IV determine the average monthly temperature for each of the two hottest months of the year and each of the two coldest months of the year in Seattle for the period 1900–1923 inclusive.

11. Table V shows the number of heads and tails that fell in each of 500 throws of seven dimes. Is it easy to determine from this table what number of heads occurred most often?

12. If the dimes of Ex. 11 had been thrown 5,000 times, would you expect the same result as to the number of heads occurring most often?

13. If the dimes of Ex. 11 had been thrown only 10 times, would you expect the same result as to number of heads occurring most often?

14. Toss 10 coins 500 times, recording the number of heads and tails at each throw, and forming a table similar to Table V.

15. Suppose that at the weather station for some locality we can obtain the daily precipitation for each day for 50 years. Suggest methods of selecting a sample of some 500 to 600 daily records from which the average daily precipitation over the entire 50 years may be obtained with a fair degree of accuracy.

16. A 100-ft. steel tape was used in measuring a distance, which was found to be 1,243.7 ft. It was afterwards found that the tape was a hundredth of a foot too long. What correction should be made in the measured distance?

CHAPTER II

TABULATION, FREQUENCY DISTRIBUTION

1. In order to compare two or more variables or to study the changes in a single variable, proper tabulation is of prime importance. This is not always an easy problem. The easiest way to tabulate a set of observations is to record each one as it is made, making a column of values in order of occurrence. Such a table is that of monthly rainfall, in Table III, in the first chapter. If it is desired to show which are the rainiest months of the year and which the driest or what amount of monthly rainfall occurs most often, the data should be tabulated in different form so as to bring out the desired information. For other characteristics it might be desirable to use a still different form of table.

2. Not many hard-and-fast rules can be given as to the form a table shall take. Experience is the one thing which gives greatest skill in making tabulations. This much may be said; there is usually some one best form to use in order to bring out the characteristics to be studied or to show the comparisons desired. It is the business of the tabulator to find this best form and use it. Frequently this is not an easy matter.

3. Best Form.—First the problem to be solved, the purpose to be served by the tabulation, should be formulated and written down. Examine the data to see if they furnish the necessary information. If not, further information must be obtained before tabulation can proceed.

Next decide whether the data shall all be placed in one table or in two or more. A large mass of data may make a single table too unwieldy. If possible, it should be easy for the eye to follow from one entry to another with which it is to be compared, or to follow through a series of entries related to each other. It may be that in order to do this data will have to be split up into several tables. Usually it is unwise to try to make one table serve for the solution of two or more distinct problems. If two series are to be compared, it is desirable to place them in adjacent columns.

The main headings, subheadings, and number of columns must be decided upon. By a proper use of main headings and subheadings, space may be saved and those things to be compared brought near to each other. Thus a larger amount of data may frequently be brought into a single table than could otherwise be done. This process can be carried too far and the table be made too large and confusing. The judgment of the tabulator and the purpose to be served will have to govern the matter.

If several different forms of tabulation are possible, it is well to sketch a form which seems to serve the purpose. Then see if this form gives a ready solution of the problem. It may soon appear that another form will make the solution of the problem easier, bring out the meaning of the figures more clearly, and make the table more easily interpreted by the reader. The judgment of the tabulator should soon tell him when the best form has been found.

4. Placing on Sheet.—Having decided on the best form to use, a sufficient number of figures should be entered to make it possible to adjust the column widths so that the table may fit the page on which it is to be finally printed. If possible adjust the table so it may be placed upright on the page with the top of the table at the top of the page. If it is necessary to make the table the other way of the sheet, place it so that a quarter right-hand (clockwise) turn will make the table right side up. This places the top of the table at the left-hand edge of the sheet. In the same way, if a column heading cannot be written across the column, it should be placed so that a quarter right-hand turn will make the printing right side up.

5. Title and Headings.—The title of the table should be as concise as possible. It must, however, make perfectly clear what material is shown. It should be unambiguous. The same remarks apply to column headings. Somewhere, either in the title or column heading, the unit of measurement used should be stated. Footnotes may be used to show the source of data or to make clear the meaning or content of any column or of the whole table. Avoid too many footnotes. It is better that each heading be completely self-explanatory if possible.

It is best that titles, column headings, and figures be all printed by hand. Ordinary script is not permissible. Different size letters can then be used according to the importance of the heading. Headings of the same degree of importance should be printed

all in the same size of letter. In some cases typewriting may be used. If so, a typewriter must be chosen that gives neat, accurate, uniform work. If the tables are to be reproduced by some photographic printing process, a typewriter ink must be used that will photograph.

6. Neatness and accuracy are prime essentials. Failure in either respect makes the work of little value. A single mistake makes the entire tabulation untrustworthy. Transcription of figures and all computations should be checked and rechecked.

7. Illustrations.—The following tables illustrate some of the points mentioned above. Suppose the data are the population and area of each state of the United States. If the purpose is merely to record the facts so that they may be available for reference, the states may be arranged in alphabetical order, as shown in Table VI.

Table VI.—Population and Area of the United States
Census of 1920

State	Population	Area, in square miles
Alabama	2,348,174	51,998
Arizona	334,162	113,956
Arkansas	1,752,204	53,335
California	3,426,861	158,297
Colorado	939,629	103,948
Connecticut	1,380,631	4,965
etc.	etc.	etc.

The purpose of this table would be served as well if the last two columns were interchanged.

If it is desired to compare poluations, to show which states have the large populations and which the small, the states should be arranged in column so that the column of populations will give the values in order of size. It may begin with the largest and proceed to the smallest or vice versa. The areas do not enter into consideration and need not be tabulated. Table VII shows the form, beginning with the state of largest population.

If it is desired to show populations by geographic distribution, the column of states may begin with those states of the New England division, then the Middle Atlantic division, and so on

across the country to the Pacific division. It will be well to show the total for each geographic division. Table VIII shows the form.

TABLE VII.—POPULATION OF THE UNITED STATES
Census of 1920

State	Population
New York	10,385,227
Pennsylvania	8,720,017
Illinois	6,485,280
Ohio	5,759,394
Texas	4,663,228
Massachusetts	3,852,356
etc.	etc.

Similar tables classifying areas only, instead of populations, may be desired. In such case population figures would not be tabulated.

TABLE VIII.—POPULATION OF THE UNITED STATES BY GEOGRAPHIC DISTRIBUTION

Census of 1920

State	Population
Maine	768,014
New Hampshire	443,083
Vermont	352,428
Massachusetts	3,852,356
Connecticut	1,380,631
Rhode Island	604,397
New England	7,400,909
etc.	etc.

8. Derived Data.—Other tables would be derived for problems involving both population and area. For instance, suppose it is desired to make a study of density of population, the population per square mile. The densities may be derived for

each state from the original population and area data. If the table is merely for reference, the states will be tabulated in alphabetic order as was done in Table VI. The states may be arranged by geographic divisions, according to numerical value of the density, by size of population, or by magnitude of area, according to what particular feature it is desired to bring out.

Table IX shows density of population with the states arranged according to size of population.

TABLE IX.—POPULATION, AREA, AND DENSITY OF POPULATION OF THE UNITED STATES

Census of 1920

State	Population	Area in square miles	Density of population, (number per square mile)
New York	10,385,227	49,204	211.1
Pennsylvania	8,720,017	45,126	193.5
Illinois	6,485,280	56,665	114.4
Ohio	5,759,394	41,040	140.3
Texas	4,663,228	265,896	17.5
Massachusetts	3,852,356	8,266	466.0
etc.	etc.	etc.	etc.

The contrast in density for Texas and Massachusetts is very noticeable, especially since Texas has a greater population than has Massachusetts. The reason is at once seen by examining the column of areas, Texas has more than 30 times the area of Massachusetts.

9. Time and Place Variation.—Population, area, density, the variables in Tables VI to IX, vary from one place to another place, time remaining constant. The same elements, population, area, density, for the United States or for any one state might have been tabulated for each census year. The variables then vary from one time to another with place remaining constant. One of these elements, say population, may be tabulated in one table as both time or historical variable and as place variable. Names of the states would be placed in column, and a column made for each date. The form is shown in Table X.

TABLE X.—POPULATION OF THE UNITED STATES AT EACH CENSUS FROM 1900 TO 1920

Source: Report of the Bureau of the Census

State	1900	1910	1920
Alabama	1,828,697	2,138,093	2,348,174
Arizona	122,931	204,354	334,162
Arkansas	1,311,564	1,574,449	1,752,204
California	1,485,053	2,377,549	3,426,861
Colorado	539,700	799,024	939,629
Connecticut	908,420	1,114,756	1,380,631
etc.	etc.	etc.	etc.

All three elements might be thus tabulated as both historical variable and place variable. In this case each census date would constitute a subheading with three columns under it, one for each of three elements, population, area, and density. The table is now likely to become large and unwieldy. If, however, it is desirable to have the data tabulated in this form, the sheet may be made large and then folded in to fit the page size of the report. Many excellent examples of tabulation of place variables, time variables, and combinations of both may be found in the "Statistical Abstract of the United States," Department of Commerce, and in many other publications from the Government Printing Office.

10. Quality Variation.—Besides the time variable and the place variable, there is what may be called the quality variable that is sometimes studied. The variable is classified according to the presence or absence of certain characteristics. Thus a certain population might be studied as to the association of blindness and deaf-mutism. The tabulation would take the form of Table XI.

TABLE XI.—BLINDNESS AND DEAF-MUTISM OF 1,163 PAUPERS

	Blind	Not blind	Total
Deaf-mute	147	12	159
Not deaf-mute	21	983	1,004
Total	168	995	1,163

Such studies lead to the theory of attributes or theory of contingency. This theory is well treated in Yule's "Introduction to the Theory of Statistics," but there will be no occasion to take it up in this book.

11. Frequency Tables.—Another very important form of tabulation is that of the frequency table, which has been defined in Chap. I. This will now be studied in more detail.

12. Classifications.—Turn to Table 1 of weights and heights of university freshmen and proceed to make a frequency table of weights containing about 30 classes. By examination of the weights, it is seen that they run from 100 to 216 lb., a range of 217 − 100 = 117 lb. Thirty is contained in 117 between three and four times, nearly four times. Use a class interval of 4 lb. The first class, then, may contain all those weights from 97 to 100 lb., inclusive, 98 to 101, 99 to 102, or 100 to 103. Considerations taken up later would determine which to use. In the table as formed, the first class includes all those weights from 97 to 100 lb., inclusive, and so on to the thirtieth class, which contains all weights from 213 to 216 lb. inclusive.

13. The finished table, Table XII, has three columns. The first column contains the weight classes, or size of items, designated by the letter X. The second column displays the frequencies, or number of items in each class, designated by the letter f. These two columns are made up directly from the original measurements as recorded in Table I. The third column, cumulative frequency, is obtained from the second column by successive additions of frequencies so that the last entry in this column equals the total frequency.

14. Class Limits and Class Boundaries.—The pairs of numbers written in the column of classes are the *upper* and *lower class limits.* Halfway between the upper limit of one class and the lower limit of the next class is a *class boundary.* The *upper boundary* of one class is the *lower boundary* of the next class. For instance, the upper boundary of the third class in Table XII is 108.5, which is the lower boundary of the fourth class. The lower boundary of the first class is 96.5 and the upper boundary of the last class is 216.5.

When the men were weighed, their weights were recorded to the nearest pound, though the scales might have been read to a fraction of an ounce. Thus all weights more than 103.5 and less than 104.5 were recorded as 104. The third class in Table XII

TABLE XII.—WEIGHTS OF FRESHMAN MEN

In pounds

Weight classes X	Frequency f	Cumulative frequency
97–100	1	1
101–104	2	3
105–108	4	7
109–112	15	22
113–116	12	34
117–120	22	56
121–124	34	90
125–128	44	134
129–132	62	196
133–136	57	253
137–140	56	309
141–144	61	370
145–148	55	425
149–152	42	467
153–156	40	507
157–160	25	532
161–164	30	562
165–168	14	576
169–172	15	591
173–176	13	604
177–180	10	614
181–184	4	618
185–188	2	620
189–192	2	622
193–196	3	625
197–200	1	626
201–204	0	626
205–208	2	628
209–212	0	628
213–216	1	629
Total................	629	

then includes all those weights which were more than 104.5 and less than 108.5, the class boundaries. The recorded weights in this class were only from 105 to 108, inclusive, the class limits.

The difference between the upper limit of one class and the lower limit of the next class is always the smallest division in the measurements used, in this case, 1 lb. The distinction between *class boundaries* and *class limits* is an important one in plotting graphs and in determining various division points.

If, in the process of weighing, there happened to be several weights that seemed to be just 103.5, half of them may be recorded as 104 and half as 103. Or, if there is an odd number, the computers' rule may be followed for the odd one which drops the 0.5 or adds on another 0.5 according to which gives an even number. In the long run, this rule results in dropping 0.5 about the same number of times that an extra 0.5 is added on, thus making the errors compensating. Or, a coin may be flipped to make the decision. The making of the frequency table is not concerned with this matter, since the measurements as recorded are the ones used.

15. Forming the Table.—Having decided on the class interval and the limits of the first class, complete the first column showing the class limits for all classes in order. Go through Table I of weights and check off each weight by a score mark opposite the class into which it falls. When this is done, add up the score marks for each class and enter the sum in the frequency column opposite the class to which it belongs. Total the frequencies. Form the cumulative frequency column. Check the last entry on the total of frequencies, and check both on the total number of individuals measured in Table I.

16. Distribution of Frequencies.—Study the frequency column. What do you note in regard to change of frequency? Do heavy-weights, light-weights, or middle-weights occur most frequently? Could this have been readily seen from Table I? Was it to have been expected? Is there regularity in change of frequency?

17. Effect of Change of Class Interval.—In the same manner that Table XII was constructed, another frequency table, of the same data, is made with a larger class interval giving fewer classes. This table, Table XIII, has eleven classes with a class interval of 11. As in Table XII, it is seen in this table that light-weights and heavy-weights occur less frequently than middle-weights. In Table XII it should be noticed that the frequencies are sometimes increasing and sometimes decreasing, but on the whole increasing up to about the middle of the table. Then they are

sometimes decreasing and sometimes increasing, but on the whole decreasing to the end of the table. In Table XIII, one of the most noticeable things is that these irregularities have been wiped out. The frequencies are everywhere increasing up to a maximum in the fourth class, and then everywhere decreasing to the end of the table. Increasing the class interval tends to smooth out irregularities in variation of frequencies. This fact will be very noticeable when graphs of these two sets of frequencies are made in a later chapter. Usually it is permissible to increase the class interval until a relatively smooth variation of frequencies is obtained. In some cases, however, certain considerations may demand the retention of the original measurements. Increasing the class interval should not be carried too far. Manifestly a class interval large enough to give only two classes would tell very little with regard to distribution of frequencies.

18. Frequency Table of Heights.—Table XIV is a frequency table of the heights in inches of the 629 freshmen given in Table I whose weights have just been tabulated. The heights were recorded to the nearest half-inch. The range was from 60.5 to 75.0 in., or 15.0 in., which is 30 units of measurement in half-inches. It being desired in a later chapter to compare the distribution of heights with that of weights, it is convenient to have about the same number of classes in Table XIV as in Table XIII. A class interval of 1.5 in. gives 11 classes, the same number as in Table XIII.

TABLE XIII.—WEIGHTS OF FRESHMAN MEN
In pounds

Weight classes, X	Frequency, f	Cumulative frequency
100–110	15	15
111–121	43	58
122–132	138	196
133–143	162	358
144–154	129	487
155–165	82	569
166–176	35	604
177–187	16	620
188–198	5	625
199–209	3	628
210–220	1	629
Total....................	**629**	

TABLE XIV.—HEIGHTS OF FRESHMAN MEN
In inches

Height classes, X	Frequency, f	Cumulative frequency
60.0–61.0	3	3
61.5–62.5	8	11
63.0–64.0	33	44
64.5–65.5	51	95
66.0–67.0	115	210
67.5–68.5	156	366
69.0–70.0	148	514
70.5–71.5	64	578
72.0–73.0	43	621
73.5–74.5	5	626
75.0–76.0	3	629
Total	629	

19. Comparison of Frequency Distributions.—Examination of the frequency column shows an increase of frequencies with increase of height up to a certain point and then a decrease to the end of the table. This was also true of the weights. It is noticeable, however, that there is greater symmetry of distribution in heights than in weights. The number of small heights is about the same as the great heights. The number of small weights is considerably more than the number of great weights. The class of greatest height frequency, 67.5 to 68.5, is at the middle of the table. The class of greatest weight frequency, 133 to 143, is nearer the beginning of the table. In the next chapter these facts will be shown at a glance of the eye by means of frequency graphs. There seems to be a greater scattering of weights as compared with the average weight than there is of heights as compared with average height. In other words, heights tend to concentrate about the average height more than weights tend to concentrate about the average weight. Numerical measures of this tendency will be devised in the chapter on dispersion.

20. Chance Distribution.—Another illustration of frequency distribution is shown by the results of pure chance taken from Table V. By tabulating number of heads as size of item and the number of times each occurred in the 500 throws as frequency, Table XV is obtained.

The similarity of the frequency distribution of Table XV to that of Table XIV is very striking. Its similarity to that of Table XII is not so great. This indicates that freshman heights follow more closely a symmetric distribution like that due to pure chance in throwing coins than do their weights. It would seem, then, that there is an intimate connection between the laws of probability governing the number of heads falling in throwing coins and the distribution of freshman heights.

TABLE XV.—NUMBER OF HEADS FALLING IN 500 RANDOM THROWS OF SEVEN DIMES

Number of heads, size of item, X	Number of times occurring, frequency, f	Per cent of total
0	1	0.2
1	31	6.2
2	76	15.2
3	146	29.2
4	144	28.8
5	76	15.2
6	23	4.6
7	3	0.6
	500	100.0

The next point of interest arising at once is to discover how closely the frequency distribution of the 500 throws of seven dimes follows the theoretical distribution called for by the laws of probability.

21. Probability Theorems.—Following are the four fundamental theorems in probability.[1]

Theorem I.—The probability that all of a set of independent events will occur is the product of the separate probabilities of the occurrence of each of the single events. This is *compound probability.*

The probability that any given event, in a set of mutually exclusive events, will happen, and all others fail, is called *partial* or *relative probability.*

Theorem II.—The probability that *any one whatever* of a set of mutually exclusive events will occur, and all the others fail, is the

[1] The proof of these laws may be found in Appendix C.

sum of the partial probabilities of the single events. This is *total probability.*

Theorem III.—If the probability that an event will occur in a single trial is p, and the probability that it will fail is q, then the probability that it will occur exactly r times, no more no less, in n trials is ${}_nC_r p^r q^{n-r}$ (($p + q$) always equals 1). ${}_nC_r$ is the number of combinations of n things taken r at a time (see Appendix B).

Theorem IV.—If the probability that an event will occur in a single trial is p, and the probability that it will fail is q, then the probability that it will occur at least r times in n trials (r times or more) is the sum of the first $(n - r + 1)$ terms in the expansion of $(p + q)^n$, that is, $p^n + {}_nC_1\, p^{n-1} q + {}_nC_2 p^{n-2} q^2 + \ldots + {}_nC_r p^r q^{n-r}$ times.

22. Theoretical Frequencies in Tossing of Coins.—In a single toss of a single dime the probability of its falling tails is $p = 1/2$. The probability of its falling heads, that is, failing to fall tails, is $q = 1/2$. In tossing one dime seven times, or seven dimes once, the probability of all tails, no heads, is $p^7 = (1/2)^7 = 1/128$, by Theorem I.

The probability that tails will occur six times and only six times, that is, that heads will occur once and only once, is ${}_7C_1 p^6 q^1 = 7(1/2)^6(1/2) = 7/128$, by Theorem III. Similarly, the probability of getting exactly 5 tails and 2 heads is

$${}_7C_2 p^5 q^2 = 21(1/2)^5(1/2)^2 = 21/128,$$

4 tails and 3 heads is

$${}_7C_3 p^4 q^3 = 35(1/2)^4(1/2)^3 = 35/128,$$

3 tails and 4 heads is

$${}_7C_4 p^3 q^4 = 35(1/2)^3(1/2)^4 = 35/128,$$

2 tails and 5 heads is

$${}_7C_5 p^2 q^5 = 21(1/2)^2(1/2)^5 = 21/128,$$

1 tail and 6 heads is

$${}_7C_6 p q^6 = 7(1/2)\,(1/2)^6 = 7/128,$$

and of getting no tails and all heads is

$$(1/2)^7 = 1/128.$$

By Theorem IV, the probability of tails occurring at least no times ($r = 0$), in other words, the probability of heads occurring at most seven times, is the sum of all of the $(n + 1) = 8$ terms of the expansion of $(p + q)^n = (1/2 + 1/2)^7$. It is certain that heads will occur seven times at most. The sum, therefore, of the eight

probabilities of the preceding paragraph should be 1, indicating certainty.

$(\frac{1}{2} + \frac{1}{2})^7 = (\frac{1}{2})^7 + 7(\frac{1}{2})^6(\frac{1}{2}) + 21(\frac{1}{2})^5(\frac{1}{2})^2 + 35(\frac{1}{2})^4(\frac{1}{2})^3$
$+ 35(\frac{1}{2})^3(\frac{1}{2})^4 + 21(\frac{1}{2})^2(\frac{1}{2})^5 + 7(\frac{1}{2})(\frac{1}{2})^6 + (\frac{1}{2})^7$
$= \frac{1}{128} + \frac{7}{128} + \frac{21}{128} + \frac{35}{128} + \frac{35}{128} + \frac{21}{128} + \frac{7}{128} + \frac{1}{128}$
$= \frac{128}{128} = 1.$

Tabulating these separate probabilities Table XVI results.

TABLE XVI.—PROBABILITIES OF OCCURRENCE OF DIFFERENT COMBINATIONS OF HEADS AND TAILS IN SEVEN THROWS OF A COIN OR IN ONE THROW OF SEVEN COINS

Heads	Tails	Probability of occurrence	Per cent of total
0	7	$\frac{1}{128}$	0.78
1	6	$\frac{7}{128}$	5.47
2	5	$\frac{21}{128}$	16.41
3	4	$\frac{35}{128}$	27.34
4	3	$\frac{35}{128}$	27.34
5	2	$\frac{21}{128}$	16.41
6	1	$\frac{7}{128}$	5.47
7	0	$\frac{1}{128}$	0.78
Total . .	. . .	. . . . $\frac{128}{128} = 1$	100.00

These eight probabilities are then proportional, respectively, to the numbers 1, 7, 21, 35, 35, 21, 7, 1, or the percentages of the last column of Table XVI. In a *very large* number of trials the theoretical distribution of number of heads calls for frequencies proportional to these numbers.

23. Comparison with Actual Trial.—If, for example, the 500 trials had followed the law of probabilities as closely as possible, the number of times 0 heads, seven tails occurred would have been 0.78 per cent of 500, or 4. The number of times one head, six tails occurred would have been 5.47 per cent of 500, or 27. Compute the theoretical frequency for each of the eight possible combinations of heads and tails. Table XVII displays the theoretical frequency f', the actual frequency f that occurred in throwing seven dimes 500 times; and the correction to be added to f to give f'.

While 500 seems like a fairly large number, it is by no means what could be called a very large number. The results of the 500 trials are fairly close to the theoretical results called for from a very large number of trials, as shown in Table XVII.

TABLE XVII.—COMPARISON OF THEORETICAL FREQUENCY WITH ACTUAL FREQUENCY FROM THROWING SEVEN DIMES 500 TIMES

Number of heads X	Theoretical frequency f'	Actual frequency (from Table XV) f	Correction to be added to f to give f'
0	4	1	+3
1	27	31	−4
2	82	76	+6
3	137	146	−9
4	137	144	−7
5	82	76	+6
6	27	23	+4
7	4	3	+1
Totals.	500	500	20–20

24. Binomial Distribution.—The theoretical distribution of frequencies derived from the expansion of $(\frac{1}{2} + \frac{1}{2})^7$ is perfectly symmetrical. There are relatively few of the extreme values of the variable, and relatively many of the medium-sized values, the maximum number occurring at exactly the middle of the table of frequencies. This distribution of frequencies is called the *symmetric binomial distribution.* It will be further discussed in the next chapter. If $p = \frac{1}{3}$ and $q = \frac{2}{3}$, and the frequencies of a distribution are proportional to the terms of the expansion of $(p + q)^n = (\frac{1}{3} + \frac{2}{3})^n$, the distribution of frequencies will no longer be symmetrical. This will always be true if $p \neq q$, and n is not too large.

25. Comparison with Freshman Heights.—Now examine the distribution of freshman heights as displayed in Table XIV. There being 11 classes, assume 11 different possible heights, namely, the midpoints of the classes, constituting a discrete series. Each differs from the next by one class interval. This class interval, 1.5 in., may be regarded as the unit of measure- [illegible] on,

the frequencies will be proportional to the 11 terms of the expansion of $(\frac{1}{2} + \frac{1}{2})^{10}$.

$$\left(\frac{1}{2}+\frac{1}{2}\right)^{10} = \left(\frac{1}{2}\right)^{10} + 10\left(\frac{1}{2}\right)^{9}\left(\frac{1}{2}\right) + 45\left(\frac{1}{2}\right)^{8}\left(\frac{1}{2}\right)^{2} + 120\left(\frac{1}{2}\right)^{7}\left(\frac{1}{2}\right)^{3}$$
$$+ 210\left(\frac{1}{2}\right)^{6}\left(\frac{1}{2}\right)^{4} + 252\left(\frac{1}{2}\right)^{5}\left(\frac{1}{2}\right)^{5} + 210\left(\frac{1}{2}\right)^{4}\left(\frac{1}{2}\right)^{6}$$
$$+ 120\left(\frac{1}{2}\right)^{3}\left(\frac{1}{2}\right)^{7} + 45\left(\frac{1}{2}\right)^{2}\left(\frac{1}{2}\right)^{8} + 10\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)^{9} + \left(\frac{1}{2}\right)^{10}$$
$$= \frac{1}{1{,}024} + \frac{10}{1{,}024} + \frac{45}{1{,}024} + \frac{120}{1{,}024} + \frac{210}{1{,}024} + \frac{252}{1{,}024} + \frac{210}{1{,}024}$$
$$+ \frac{120}{1{,}024} + \frac{45}{1{,}024} + \frac{10}{1{,}024} + \frac{1}{1{,}024} = \frac{1{,}024}{1{,}024} = 1.$$

The frequencies of the symmetric binomial of distribution will, therefore, be proportional to the numbers in Table XVIII.

TABLE XVIII

Number	Per cent of 1,024
1	0.1
10	1.0
45	4.4
120	11.7
210	20.5
252	24.6
210	20.5
120	11.7
45	4.4
10	1.0
1	0.1
Totals......... 1,024	100.0

The theoretical frequency to the nearest integer for the first class is 0.1 per cent of 629, or 1; for the second class is 1.0 per cent of 629, or 6; for the third class is 4.4 per cent of 629, or 28; and so on.

Table XIX shows the values of f', f, and the corrections, as was shown in Table XVII for the results of throwing the coins.

From an examination of Table XIX it is seen that the distribution of heights of the 629 freshmen is fairly close to the

TABLE XIX.—COMPARISON OF FREQUENCY DISTRIBUTION OF FRESHMAN HEIGHTS AND SYMMETRIC BINOMIAL DISTRIBUTION

Midpoints of Classes X	Theoretical frequency f'	Frequency of freshman heights f	Correction to be added to f to give f'
60.5	1	3	− 2
62.0	6	8	− 2
63.5	28	33	− 5
65.0	73	51	+22
66.5	129	115	+14
68.0	155	156	− 1
69.5	129	148	−19
71.0	73	64	+ 9
72.5	28	43	−15
74.0	6	5	+ 1
75.5	1	3	− 2
Totals.	629	629	46–46

theoretical, chance distribution resulting from throwing 10 coins a very large number of times.

26. Non-symmetric Freshman Weights.—As was previously noted, the distribution of freshman weights was not symmetrical. Perhaps values of p and q $(p \neq q)$ could be found such that the terms of the binomial expansion of $(p+q)^n$ would be closely proportional to the frequencies of freshman weights. This involves a problem in *curve fitting* which will not be treated here.

27. Experience shows that very many of the variables dealt with by the statistician have frequency distributions approximating the symmetric binomial distribution. Certain statistical constants are derived on the assumption of such distribution.

28. Double Entry.—If two variable characteristics of the same set of things are measured, the values of the variables may be tabulated so as to display both together in such a manner as to show whether a change in one variable tends to accompany a corresponding change in the other or not. The sheet is ruled both ways into *cells* making a table of double entry. One variable is tabulated horizontally and the other vertically. Table II of Ex. 2, Chap. I, is a table of this kind. The method can best be illustrated by an example. Suppose the lengths and breadths

TABLE XX.—LENGTHS AND BREADTHS OF 60 LEAVES SELECTED AT RANDOM FROM A SINGLE TREE

All measures are to the nearest millimeter

Length l	Breadth b	Length l	Breadth b	Length l	Breadth b
21	16	33	27	40	26
21	18	33	28	41	27
23	19	34	28	41	25
24	22	34	31	41	25
25	19	34	30	42	30
25	22	35	25	42	29
26	21	35	23	43	29
26	21	35	23	43	28
27	24	35	25	44	31
27	25	36	27	44	30
27	25	36	24	44	31
28	27	37	26	45	25
30	20	37	28	45	27
30	22	37	26	45	27
30	22	37	28	46	26
31	23	38	30	46	28
31	25	38	29	47	30
31	25	38	29	48	33
31	25	39	32	50	32
33	26	40	22	53	34

of 60 leaves, selected at random from a tree, are measured and tabulated, the results being shown in Table XX. They are tabulated according to length. It may be noticed that with increase of length there is not always an increase of breadth, though, in general, the leaves get wider as they get longer.

29. Construction of Table.—Draw up a diagram of cells for a table as suggested above, tabulating lengths vertically and breadths horizontally. This provides a column of cells for each possible breadth from the smallest that occurs to the largest. Similarly, there is a line of cells for each possible length from the smallest that occurs to the largest. Call this Table XXI. Using Table XX for each pair of length and breadth (l, b), make a check mark in Table XXI in that cell which is in the line l and in column b. The method is readily seen by going through Table XX and following the results in Table XXI.

TABLE XXI.—LENGTHS AND BREADTHS OF 60 LEAVES SELECTED AT RANDOM FROM A SINGLE TREE

All measurements to the nearest millimeter

l \ b	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	Totals
21	′		′																	2
22																				0
23				′																1
24							′													1
25				′			′													2
26						″														2
27									′	″										3
28												′								1
29																				0
30					′		″													3
31								′		‴										4
32																				0
33											′	′	′							3
34													′		′	′				3
35								″		″										4
36									′			′								2
37											″		″							4
38														″	′					3
39																	′			1
40							′				′									2
41										″		′								3
42														′	′					2
43													′	′						2
44															′	″				3
45										′		″								3
46											′		′							2
47															′					1
48																		′		1
49																				0
50																	′			1
51																				0
52																				0
53																			′	1
Totals	1	0	1	2	1	2	5	3	2	10	5	6	6	4	5	3	2	1	1	60

TABLE XXII.—LENGTHS AND BREADTHS OF 60 LEAVES SELECTED AT RANDOM FROM A SINGLE TREE

All measurements to the nearest millimeter

l \ b	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	Totals
21	1		1																	2
22																				0
23				1																1
24							1													1
25				1			1													2
26						2														2
27									1	2										3
28												1								1
29																				0
30					1		2													3
31								1		3										4
32																				0
33											1	1	1							3
34													1		1	1				3
35								2		2										4
36									1			1								2
37											2		2							4
38														2	1					3
39																	1			1
40							1				1									2
41										2		1								3
42														1	1					2
43													1	1						2
44															1	2				3
45										1		2								3
46											1		1							2
47															1					1
48																		1		1
49																				0
50																	1			1
51																				0
52																				0
53																			1	1
Totals	1	0	1	2	1	2	5	3	2	10	5	6	6	4	5	3	2	1	1	60

Table XXI might have been prepared directly when the original measurements were made, not making Table XX at all.

From this table prepare another table exactly like it, except that the entry in each cell is the sum of the score marks in Table XXI. This is Table XXII.

30. Scatter Diagram.—The entries in Table XXII tend to follow the diagonal down to the right, there being no entries in the lower left-hand part, nor in the upper right-hand part of the table. These facts show that short leaves are in general narrow and long leaves are wide, and that no long leaves are short or wide leaves long. It is thus seen at a glance of the eye that there is a strong tendency for variable *b*, breadths, to increase with increase of variable *l*, lengths. If the entries had followed the other diagonal, it would have shown that as one variable increased, the other decreased.

In the chapter on correlation this table will be used to illustrate the method of getting a measure of the tendency of one variable to change along with another. Such a table as Table XXII is called a *scatter diagram* or *correlation table.*

31. Frequency Tables.—A frequency table for each variable can now be easily drawn up from Table XXII. Having decided on the class limits for each frequency table, draw, in Table XXII, heavy vertical lines at the class boundaries of the horizontal variable, and heavy horizontal lines at the class boundaries of the vertical variable. The sum of the entries between heavy lines in the vertical column of totals gives the frequency column for the vertical variable. In the same way, the sum of entries between heavy lines in the horizontal line of totals gives the frequency column for the horizontal variable. Table XXIII and Table XXIV are thus derived from Table XXII.

TABLE XXIII.—FREQUENCY TABLE OF LEAF LENGTHS

Classes (size of item in millimeters) l	Frequency f
20–24	4
25–29	8
30–34	13
35–39	14
40–44	12
45–49	7
50–54	2
Total..............................	$N = 60$

TABLE XXIV.—FREQUENCY TABLE OF LEAF BREADTHS

Classes (size of item in millimeters) b	Frequency f
15–17	1
18–20	4
21–23	10
24–26	17
27–29	16
30–32	10
33–35	2
Total	$N = 60$

Exercises

1. Taking Table II, draw horizontal and vertical lines separating the classes of Table XIII and Table XIV respectively. Check the frequency columns of Tables XIII and XIV.

2. Make a frequency table of mean monthly temperatures at Seattle as recorded in Table IV. Use a 5° class interval. Let the class limits for the first class be 30.0–34.9°.

3. Make a frequency table of monthly precipitation in inches at Seattle as recorded in Table III. Use a 1-in. class interval. Let the class limits for the first class be 0.00–0.99 in.

4. From Table V construct a frequency table for all possible numbers of heads that could occur. Does the result check Table XV?

5. Make a scatter diagram of Seattle mean monthly temperature and precipitation as shown in Table III and Table IV. Make the horizontal variable precipitation in inches, and the vertical variable mean monthly temperature in degrees Fahrenheit.

6. Make a frequency table of the mean monthly temperature in Seattle as shown in Table IV. Use a 2° class interval.

7. Using Table XXII and a class interval of 5 mm., construct five different frequency tables of leaf lengths. One of these is the same as Table XXIII. If it is desirable to have the frequencies increase rather steadily up to a maximum and then rather steadily decrease, which of the five tables is best to use?

8. Using Table XXII and a class interval of 3 mm., construct three different frequency tables of leaf breadths. If it is desirable to have the frequencies increase rather steadily up to a maximum and then rather steadily decrease, which of the three tables is best to use?

9. Using Table XXII, construct two frequency tables of leaf lengths, one with a class interval of 2 mm., and the other with a class interval of 6 mm. In each table let the lower limit of the first class be 20 mm. In which table is there greater regularity of increase and decrease of frequencies?

10. From the table constructed in Ex. 14, Chap. I, make a frequency table of number of heads similar to Table XV.

11. Expand $(\frac{1}{2} + \frac{1}{2})^{10}$ and so determine the theoretical frequencies for the number of heads in 500 tosses of 10 coins each.

12. Make a table like Table XVII, comparing the actual results of Ex. 10 with the theoretical results of Ex. 11.

13. Obtain from the 1920 census reports, or elsewhere, the population of each of the states and their areas. Tabulate the data with reference to size of population, as in Table VII.

14. Obtain the area of each of the states and tabulate the data according to size of the states.

15. From the data of the tables in Exs. 13 and 14 complete Table IX for the entire country.

16. Rearrange the table of Ex. 15 according to density of population, beginning with the state having the greatest density.

17. From census reports, or elsewhere, obtain the population of each of the states for 1910 and for 1920. Compute the per cent increase for each state. Tabulate with respect to per cent increase, beginning with the state having the greatest per cent increase.

18. From a tree select a single branch containing about 300 full-grown leaves. Try to select a branch so that the leaves on it are representative of the entire universe of leaves on the tree. Measure the length and breadth of each leaf, recording in the manner of Table XXI. From this record complete a table in the form of Table XXII.

19. From the table of Ex. 18 construct a frequency table of leaf lengths having about 10 or 12 classes.

20. From the table of Ex. 18 construct a frequency table of leaf breadths having about 10 or 12 classes.

CHAPTER III

GRAPHICAL REPRESENTATION, FREQUENCY GRAPHS

A graphical representation of statistical data is variously called *graph, chart,* or *diagram*. There are many forms. Several very good books on this subject have recently been published.[1] Some of the commoner forms will be considered in this chapter.

1. Object of a Graph or Chart.—The object of a graph or chart is to show to the eye the changes of a variable so as readily to compare one variable with another. By means of the graph or chart a comprehensive view of quantitative data is obtained which is much more easily remembered than are the tabulated numerical values. For this reason graphs are very valuable in presenting statistical data to the general public in lectures, publications, window displays, or other exhibits. Frequently relations are easily discovered by graphs which would not be readily seen from the figures themselves.

This makes the graph of value to the business executive, the engineer, and the research student in nearly all lines. Study of graphs of time variables reveals what has happened in the past and frequently enables one to predict what will occur in the future if the same influences are continued. This is of great value to the financier, the economist, the executive, and many others. The remarks in Chap. I in regard to title, arrangement, and accuracy should be reviewed before making a finished graph.

2. Bar Diagram.—The simplest form of statistical graph is the bar diagram. The numbers in a table of values of a variable are represented by horizontal or by vertical lines or bars whose lengths, measured to some scale, are proportional to the numbers

[1] "Charts and Graphs," KARL G. KARSTEN, 724 pp., published by Prentice-Hall. "How to Make and Use Graphic Charts," 539 pp., and "Graphic Charts in Business," 250 pp., both by ALLAN C. HASKELL, and published by Codex Book Company. "Graphical Methods," WILLIAM C. MARSHALL, 253 pp., published by McGraw-Hill Book Company, Inc. The last two contain bibliographies of books, pamphlets, and articles relating to the subject. "The Statistical Atlas of the United States" contains many forms not treated in this chapter.

represented. This is illustrated by the data of Table XXV, Fig. 1 and Fig. 2.

Figure 1 shows the horizontal representation of the data and Fig. 2 the vertical-bar representation. Which to use may usually be left to the judgment. It depends on the quantity of data,

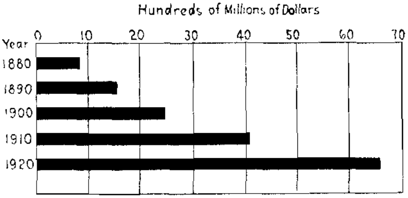


Fig. 1.—Savings banks deposits in the United States.

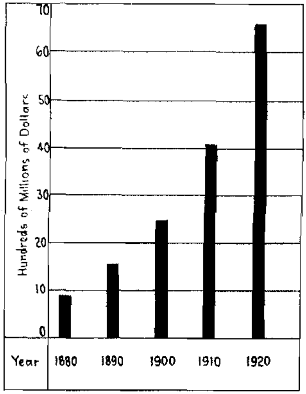


Fig. 2.—Savings banks deposits in the United States.

the scale to be used, and the size and shape of the sheet. For a simple sequence of numbers, such as in Table XXV, one way seems to show the changes in the variable about as well as the other. For public displays, horizontal bars are very frequently

TABLE XXV.—SAVINGS-BANKS DEPOSITS IN UNITED STATES

Year	Deposits
1880	$ 819,106,973
1890	1,524,844,506
1900	2,449,547,885
1910	4,070,486,246
1920	6,536,470,000

used. For office study, vertical bars, or graphical forms, to be shown later, are very common.

If the vertical bars, in this illustration, had been placed as close together as were the horizontal bars, the figure would not have been well proportioned. Spacing between the bars should be uniform, as should also be the width of bars. Width of bars, spacing between the bars, and a proper scale should be determined from a study of the data and the size and shape of the finished chart.

3. Several Variables.—Two or more variables may be represented in the same chart by representing each by bars of different color or cross-hatched in a different manner. Showing too many variables in the same chart makes the chart confusing.

4. Subdivisions.—If the variable to be represented by bars may be divided into several parts, the bar corresponding to each value of the variable may be divided into parts whose lengths

TABLE XXVI.—GRAIN CROPS OF FOUR STATES

Grain	State A		State B		State C		State D	
	Millions of bushels	Per cent of total	Millions of bushels	Per cent of total	Millions of bushels	Per cent of total	Millions of bushels	Per cent of total
Wheat	5	25	15	43	20	80	23	51
Corn	10	50	9	26	3	12	12	27
Oats	5	25	11	31	2	8	10	22
Totals	20	100	35	100	25	100	45	100

correspond to the values of the parts of the variable. For example, suppose the population of a city is 350,000, of which 290,000 are whites, 30,000 negroes, 20,000 Orientals, and 10,000

Indians. The bar representing 350,000 would be divided into parts proportional to the numbers 290,000, 30,000, 20,000, and 10,000, each part being of a different color or cross-hatched differently. The parts of all bars should be arranged in the same order. Those parts at the beginning of the bars are easily compared. Those farther on are not so easily compared. A legend should be made in a convenient place (usually the lower right-hand part of the sheet), showing the meaning of each color or

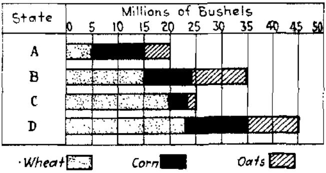


FIG. 3.—Grain crops of four states.

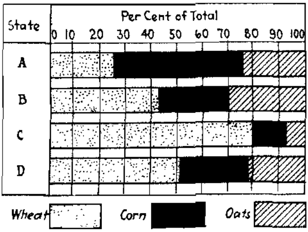


FIG. 4.—Grain crops of four states. Percentages.

cross-hatching. Figure 3 illustrates this variety of chart, the data being taken from Table XXVI.

5. One-hundred Per Cent Bars.—Sometimes it is desired to portray each part as a per cent of the whole. Then each bar is a 100 *per cent bar* and all are the same length. The parts of each bar are proportional to the corresponding percentages. Figure 4 is a percentage chart of the data shown in Fig. 3. The percentages are tabulated in Table XXVI.

6. Pie Diagrams.—Another form of graphic representation is known as the *pie diagram*. This may be used for variables that are divided into parts such as the various grain crops of differ-

ent states of Table XXVI. Using the data of this table for illustration, draw a circle for each state. The areas of these circles are made proportional to the total grain crops of the respective states. This requires the radii to be proportional to the square roots of the total crop for each of the respective states. The circumference of each circle is divided into segments proportional to the amounts of each crop in the corresponding state. Radii are drawn to the points of division. This makes

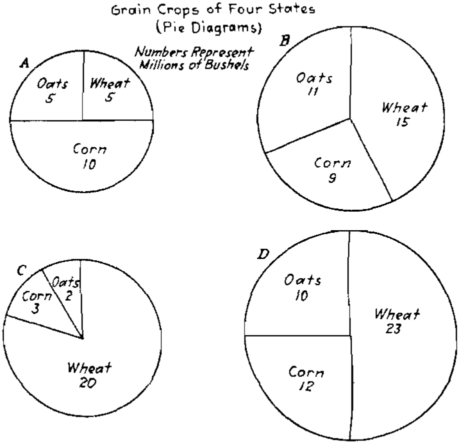


FIG. 5.

the amount of any crop proportional to the area of the corresponding circle sector. The results are shown in Fig. 5.

For practical purposes of comparison of each crop in different states, this kind of diagram is useless. For example, the corn crops of A and D are difficult to compare without referring to the numbers entered on the circle sectors, or to the length of the bar segments of Fig. 3. The radius of the corn sector in A is less than that in D, while the arc is greater. This makes it almost impossible to compare the areas by eye.

7. Percentage Pie Diagrams.—The case is not quite so bad if percentages are used. Each circle is then a 100 per cent circle and all have the same radius which is purely arbitrary in length.

The result is shown in Fig. 6. To compare percentages of different crops in the same state, or of the same crop in different states, it is only necessary to compare lengths of arcs or size of angle at the center instead of different shaped areas. Lengths of straight lines are more easily compared by the eye than lengths of curved lines or sizes of angles. So for the purposes of study in the office, or for research, the pie diagram is nearly worthless.

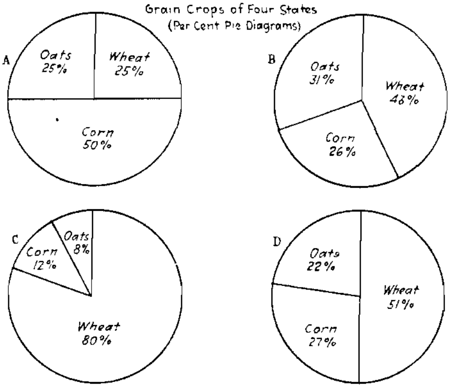


FIG. 6.

8. Use of Pie Diagrams.—The pie diagram, however, is very popular for display purposes. It catches the public eye. It appeals to the popular mind, especially in representing financial data. An example is the division of the working-man's dollar, showing what per cent goes for food, what for rent, etc. The dollar is round; so is the pie diagram. The dollar is divided into a hundred cents. The per cent pie diagram is divided into a hundred equal parts. Thus the division of the dollar is easily understood by means of this sort of diagram. If a number of charts are being displayed, it may rest the eye and hold attention if the form of chart is varied from time to time. The pie diagram may well serve this purpose.

9. Pictograms.—Sometimes resort is made to pictures for representing the various sizes of the variable, the size of the picture being proportional to the value represented. Such graphs may be called *pictograms*. As illustrative examples, may be mentioned size of armies by pictures of soldiers proportioned to the sizes of armies represented, or growth of the navy by pictures of battleships, or comparison of wheat crops of the countries of the world by varying sizes of sacks of wheat, or varying sizes of piles of sacks. Many examples could be mentioned.

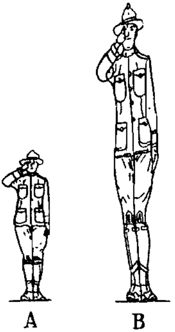


FIG. 7.

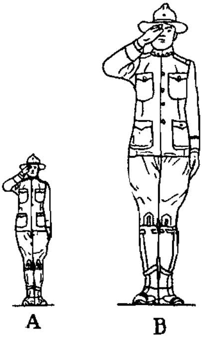


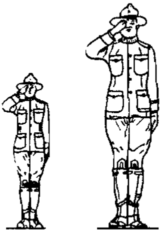
FIG. 8.

10. Methods of Representation.—As an example represent the size of an army by a picture of a soldier. Suppose army A has 500,000 men and army B has 1,000,000 men. Army B is twice as large as army A. (1) Shall the picture of soldier B be twice as tall as soldier A? (2) Or shall the picture of soldier B be twice as big (twice the area) as that of soldier A? (3) Or shall we assume soldier B to be twice as big (twice as heavy) as soldier A?

If the first method of representing the sizes of the armies by the tallness of the soldier is used, the other dimensions should strictly remain constant. Then soldier A will look very fat or soldier B will look very thin, making the pictures grotesque, as shown in Fig. 7.

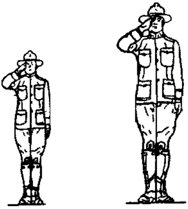
To make the pictures look well, it will be necessary to increase the width of B's picture in the same ratio as the height. The

result is shown in Fig. 8, where B is twice as tall and twice as wide as A. But now, the pictures of A and B being the same shape with the dimensions of the picture of B twice the corresponding dimensions of that of A, the area of the picture of B is four times that of A. People accustomed to comparing areas will see at a glance that the size of the picture of B is about four times that of A. Many people would interpret the picture of B as being more than twice the size of that of A but not four times as large. A smaller percentage of people, accustomed to comparing volumes, would interpret soldier B as being not only twice as high and twice as wide as soldier A but also twice as thick. In that case B would



A B

Fig. 9.



A B

Fig. 10.

be judged to be eight (2^3) times as large as A. It is very evident, then, that Fig. 8 would be deceptive to most people. Some few would say that B is about eight times as large as A. More would think that B is about four times as large as A. Still others would judge B to be about three times as large as A. In most cases, only by mental effort would the heights alone be compared and the true result be obtained.

Perhaps the best method would be to make the area of the picture of B twice that of A method (2). In this case the ratio of corresponding dimensions will be $1:\sqrt{2}$. The result is shown in Fig. 9. Of course, there is now difficulty for those people that would consider three dimensions and thus compare size or weights of the two soldiers, and for those people that would compare heights only.

If the sizes or weights of the two soldiers are compared, the corresponding dimensions of the pictures should be in ratio $1:\sqrt[3]{2}$

—method (3). Figure 10 has been made on this supposition. Probably but few people looking at this picture would interpret soldier B as being twice as large as soldier A. Nevertheless, this is the true picture if the two soldiers shown are of the same shape.

11. Use with Caution.—The conclusion may readily be drawn that the pictogram is not a desirable form of presentation of statistical data if accuracy of representation is desired. The pictogram makes a very popular form of appeal and is almost sure to attract the attention. No matter which of the above methods of construction is used, the pictogram is almost sure to be deceptive to certain classes of people. For advertising purposes, it should, therefore, be used only with great caution. A person observing pictograms should be able to decide which method of construction has been used and interpret the pictures accordingly.

12. Simplest Form Best.—Since practically everyone can, by the eye, compare areas more easily than volumes, and lengths more easily than areas, those forms of graphical representation which require the comparison of lengths only are the simplest and best to use. It happens that they are also the easiest to construct. Others may sometimes be resorted to for display purposes, to attract attention, or when the spectacular is desired. The general public is rapidly becoming accustomed to graphical data that are to be compared. This fact is eliminating the necessity of the spectacular, for gaining attention.

13. Functional Relationship.—In studying the changes of variables, it has been seen that the value of a variable changes along with the change in value of another variable. The historical or time variable changes with change of time. In a frequency distribution, the value of frequency varies with the size of the object measured. When two variables are so related that the value of one depends on the value of the other, the first variable is called a *function* of the second. The variable whose value depends on that of the other is also called the *dependent variable*, the other the *independent variable*. Thus in Table XXV the amount of savings-banks deposits may be regarded as a function of the time or date. It would then be called the dependent variable depending on time, the independent variable. The amount of deposits appears as a function of time, in this case, always increasing with increase of time. A very good example of functional relationship is the distance passed over by a falling

body in a given length of time. In this case the functional relationship may be expressed by a mathematical equation,

$$s = \frac{1}{2}gt^2.$$

Where s is the number of units of space passed over in t units of time, and g is a constant depending on gravitational attraction, and the particular units used in measuring space and time. The dependent variable, space, s, is a function of the independent variable, time, t. The form of the function is $\frac{1}{2}gt^2$, and can at once be determined for any given value of t.

14. Graph of a Function.—In the graphical representation of a function the values of the independent variable are usually represented in the horizontal direction and the values of the function, vertically. This is illustrated in Fig. 2. The independent variable, time, is plotted on a horizontal scale, and dependent variable, savings-banks deposits, on a vertical scale. In Fig. 2 *the only values of the dependent variable shown are represented* by vertical bars at 10-year time intervals, these being the values shown in Table XXV.

15. Graphical Interpolation.—For any date between the tabulated dates of Table XXV, there existed values of savings-banks deposits, but it is not known what the actual values were. In other words, for values of the independent variable between those values tabulated, there correspond values of the dependent variable. These intermediate values may be approximated by connecting the tops of the bars in Fig. 2 by straight lines. Then, at any intermediate date on the horizontal scale, go up to the straight line and read the value on the vertical scale at the left. This process of obtaining graphically approximate values of the dependent variable for values of the independent variable in between the tabulated values is called *graphical interpolation*. The method of connecting the points by straight lines, of course, assumes a uniform yearly change in savings-banks deposits during any given 10-year period.

16. Line Diagram.—It is unnecessary actually to construct the vertical bars. Instead, a single point may be plotted for each date of the table. A point for any date should be directly above the horizontal scale point for that date at a distance equal to the amount of savings-banks deposits for that date, as measured on the vertical scale. When a point has been plotted for each date given in the table, then each point may be connected with the next by a straight line. Instead of connecting the points by

straight lines, a smooth curve may be drawn passing through all the points. Figure 11 shows the points for Table XXV plotted and connected by straight lines. Such a graph is called a *line diagram*, or *peak-top diagram*. The name "peak-top" arises from the fact that a line diagram for a variable which alternately increases and decreases has somewhat of a saw-tooth appearance with peaks and low points. One author[1] calls these graphs *curves*.

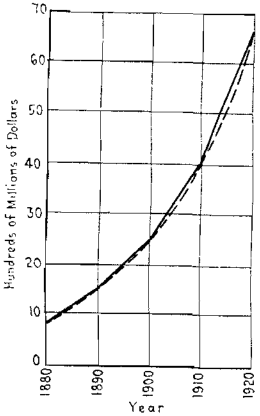


FIG. 11.—Savings banks deposits in the United States.

17. Smoothing.—The dotted line of Fig. 11 is a *smooth curve* passing through the plotted points. Interpolation may be made to the smooth curve or to the straight lines. Interpolation to the smooth curve does not imply a uniform variation between tabular dates. Neither does it imply a sudden change in variation at tabular dates as does the line diagram. If the nature of the changes in the dependent variable between tabular dates is not known, it certainly will not be justifiable to draw a smooth curve. In that case the straight lines merely serve as a guide to the eye in following from one plotted point to another and to give a general picture of the change from date to date. Interpolation

[1] KARL G. KARSTEN, "Charts and Graphs."

would not be justified unless modified by the statement that the result is true if the variation from point to point is uniform. If reasonable assumptions can be made in regard to the nature of the variation of the dependent variable between tabular entries, a smooth curve may be drawn to conform to them. Judgment and experience are the best guides to follow in smoothing. A fairly good rule in many cases is to think of the points as plotted on a level field and mark the path of a bicycle ridden through the points. There are no sudden changes in direction of the curve. In other words, it is smooth at all points. The method of smoothing the graph of a frequency distribution is somewhat different.

18. Plotting a Function.—In case of an equational relationship between the independent variable and its function, such as

$$s = \tfrac{1}{2}gt^2,$$

the graph is a determinable, mathematical curve. A great enough number of points may be determined so that the curve may be plotted with a high degree of accuracy. Using $s = \frac{1}{2}gt^2$ and supposing the conditions to be such that $g = 32$, the equation becomes $s = 16t^2$. Make a table of values of the function s (dependent variable) for a series of values of the independent variable t, Table XXVII.

TABLE XXVII

$s = 16t^2$	
t	
0	0
1	16
2	64
3	144
4	256
5	400

Choose proper horizontal and vertical scales and plot the values of the function for the series of values of $t = 0, 1, 2, 3, 4,$ and 5, as shown in Fig. 12. Where the points are as far apart as at $t = 4$, $s = 256$ and $t = 5$, $s = 400$, intermediate values of t may be chosen, s computed, and the resulting points plotted. The point at $t = 4\frac{1}{2}$, $s = 324$ is such a point. At $t = 1$, $s = 16$, the direction of the curve is changing so rapidly that intermediate points are again useful. Points at $t = \frac{1}{2}$, $s = 4$, and at $t = 1\frac{1}{2}$,

$s = 36$ will be convenient to use. Having plotted a sufficient number of points, the curve may be sketched in or drawn by help of a French curve or a flexible ruler. Since s varies not uniformly but as the square of t, the points should not be connected by straight lines. The curve is called the graph of the function $16t^2$. The process of constructing the graph is called *plotting the function*. With this curve the value of the function s may be at once read off for any value of t. Graphical interpolation may be made

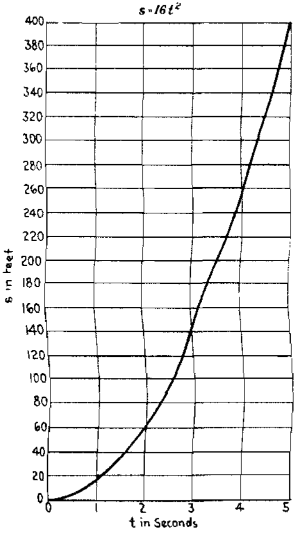


FIG. 12.

anywhere within the limits of the figure and to a degree or accuracy depending on the care taken in plotting the points, drawing the curve, and reading off the required amounts.

19. Several Dependent Variables.—With line diagrams, either peak-top or curve, two or more variables dependent on the same independent variable may be plotted above the same horizontal scale. As an example one might use deposits in savings-banks,

value of agricultural products, and number of college students, at different dates. Time would be the independent variable, common to all three dependent variables. The data, given in Table XXVIII, are graphed in Fig. 13.

TABLE XXVIII.—SAVINGS-BANKS DEPOSITS, VALUE OF WHEAT CROP, AND NUMBER OF COLLEGE STUDENTS IN UNITED STATES

Year	*Deposits, in hundreds of millions of dollars*	*Value of wheat crop, in hundreds of millions of dollars*	*Number of college students, in thousands*
1880	8.2	3 9	38.2
1890	15.2	3 2	55.7
1900	24.5	4.2	98.9
1910	40.7	5.8	163.0
1920	65.4	14.9	441.4

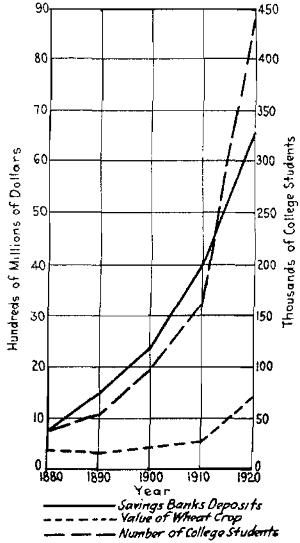


FIG. 13.—Savings banks deposits, value of wheat crop and number of college students in the United States.

For number of college students a scale different from that for money values must be used. This scale has been placed at the right. Sometimes when different scales are necessary, they are placed side by side at the left of the chart. More than two or three scales would be confusing. It is not well to graph too many variables in one chart. They become difficult to follow. It is better to group them in two or more charts. The advantage in plotting several variables in one chart is that their variations can be easily compared. In Fig. 13 it is very noticeable that the number of college students varies in about the same manner as deposits in savings banks. This at once suggests correlation between these two variables. Study this graph and see what other conclusions can be drawn.

20. Axes.—In connection with graphs of this sort, certain mathematical terms should be recalled from algebra. In Fig. 12 the horizontal scale line passes through the 0 of the vertical scale, and the vertical scale line passes through the 0 of the horizontal scale. These lines through the 0 of both scales are called *coordinate axes.* The horizontal line through the 0 of the vertical scale is called the *axis of abscissas.* The vertical line through the 0 of the horizontal scale is called the *axis of ordinates.* In Fig. 12 the axis of abscissas may also be called the *t-axis,* and the axis of ordinates the *s-axis,* named from the variables represented. The axis of abscissas is the axis of the *independent variable.* The axis of ordinates is the axis at the *dependent variable* or the *function axis.* The point of intersection of the two axes, the common zero point of both scales, is called the *origin of coordinates* or simply the *origin.* In passing, it may remarked that not in all graphs are the scale lines drawn through the origin.

21. Coordinates.—The two numbers which determine the position of a plotted point are called the *coordinates* of the point. The coordinates of a point are its distances from the axes, as measured by the horizontal and vertical scales used. The distance from the vertical axis is called the *abscissa* of the point. The distance from the horizontal axis is called the *ordinate* of the point. Points above the axis of abscissas have positive ordinates. Points below the axis of abscissas have negative ordinates. Points to the right of the axis of ordinates have positive abscissas. Points to the left of the axis of ordinates have negative abscissas. The two coordinates of a point are written together in a parenthesis with a comma between them, the abscissa first followed

by the ordinate, thus: (3, 4). The point (3, 4) is three horizontal scale units to the right of the axis of ordinates and four vertical scale units above the axis of abscissas. The point (−3, 4) is three horizontal scale units to the left of the axis of ordinates and four vertical scale units above the axis of abscissas. The points plotted in Fig. 12 as taken from Table XXVII were (0, 0), (1, 16), (2, 64), (3, 144), (4, 256), and (5, 400).

The preceding definitions are according to the usual conventions. The axes might have been assumed in other positions and positive and negative in other directions, giving another coordinate system. The usual conventions are implied unless otherwise stated.

22. x and y Independent.—Designate two variables by the letters x and y, as is commonly done in algebra. Let x be the independent variable and y the dependent variable. The horizontal axis or axis of abscissas is the x-axis. The vertical axis or axis of ordinates, the function axis, is the y-axis. The point (x, y) is any point in the plane of the axes, with abscissa equal to x and ordinate equal to y. If the values of neither x nor y are in any way dependent on the value of the other, the point (x, y) may wander all over the plane from one place to another by merely assuming arbitrary values for x and for y.

23. Straight Line through the Origin.—But suppose that x and y are subject to the condition that their ratio, y/x, must remain constant. For example, suppose $y/x = 2$. Then $y = 2x$. One is limited to consideration of that set of points such that the ordinate of each is just twice its abscissa. For $x = 0$, $y = 0$: $x = 1$, $y = 2$; $x = 2$, $y = 4$; $x = -1\frac{1}{2}$, $y = -3$, etc. For any change in x the change in y is just twice as much. In other words y changes uniformly with a uniform change in x. It follows that all the points whose coordinates satisfy the condition $y = 2x$ must lie on a straight line through the point (0, 0), the origin. Or, consider the matter in this manner. In Fig. 14 take any two points, P_1 and P_2, whose coordinates satisfy the condition $y = 2x$. Draw the abscissas OA_1, OA_2 and the ordinates A_1P_1, A_2P_2. By the given conditions

$$\frac{A_1P_1}{OA_1} = \frac{A_2P_2}{OA_2} = 2.$$

Therefore, from the similar triangles OA_1P_1 and OA_2P_2, the points O, P_1, and P_2 must lie in a straight line. Since this conclusion holds true for *any* two points P_1 and P_2 of the set of

points satisfying the condition, it holds for all points of the set. So all points, whose coordinates satisfy the condition $y = 2x$, must lie on a straight line through (0, 0). Moreover, the coordinates of any point off that straight line will not satisfy the given condition. The ratio of ordinate to abscissa will be either more or less than 2 according to the side of the line on which the point falls

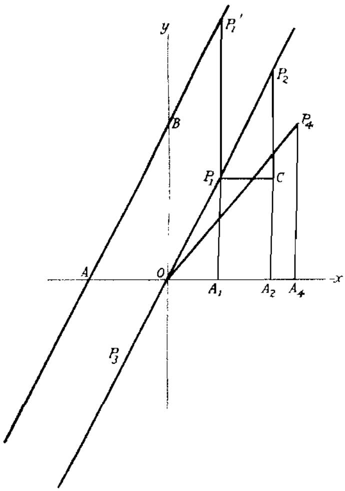


FIG. 14.

For P_4,

$$\frac{A_4P_4}{OA_4} < 2.$$

24. Locus, Slope.—The straight line P_2P_3 which contains all of the points and only those points whose coordinates satisfy the condition $y = 2x$ is called the *locus* of the equation $y = 2x$. It is the graph of the function $2x$. $y = 2x$ is called the *equation of the line* P_2P_3, the coordinates of every point of which, and only those points, satisfy the equation. The ratio $y/x = 2$ is called

the *slope* of the line. The slope may be obtained by taking any two points on the line, such as P_1 and P_2, and finding the ratio of the difference of the ordinates of these two points to the difference of their abscissas.

$$\frac{A_2P_2 - A_1P_1}{OA_2 - OA_1} = \frac{CP_2}{A_1A_2} = \frac{CP_2}{P_1C} = \text{slope}.$$

The same arguments apply to the condition $y/x = m$, any constant. So the locus of the general equation of this form $y = mx$ must be a straight line through the origin (0, 0) with slope m.

25. Straight Line Not through the Origin. Slope Form.—Take the point B (0, 3) on the y-axis. Through B draw a line parallel to P_2P_3. It is evident, at once, that the ordinate of any point P_1' on this line is three units greater than the ordinate of the point on P_2P_3 having the same abscissa. It must be true, therefore, for every point on this line, and only those points, that

$$y = 2x + 3.$$

The distance 3 from the origin where this line cuts the y-axis is called the *y-intercept.* The *x-intercept* may be found by putting 0 for y in the equation and solving for x.

$$0 = 2x + 3,$$
$$2x = -3,$$
$$x = -\tfrac{3}{2} = -1\tfrac{1}{2}.$$

The point $(-1\tfrac{1}{2}, 0)$ is on the x-axis and on the line. Why?

The line $P_1'A$ is the locus of the equation $y = 2x + 3$, and the graph of the function

$$2x + 3.$$

$y = 2x + 3$ is the equation of the line $P_1'A$.

The same line of reasoning shows that if the line is moved parallel to itself until it cuts the y-axis at a distance b from the origin, its equation will become

$$y = mx + b. \qquad (1)$$

This is called the *slope form* of the equation of a straight line.

The function $mx + b$, where m and b are constant and x variable, is called a *linear function* of x, because its graph is a straight line.

26. First-degree Equation.—Any first-degree equation in x and y can be reduced to the form

$$y = mx + b$$

simply by solving for y. Take the general form of a first-degree equation in x and y,

$$Ax + By + C = 0. \tag{2}$$

Solving for y gives

$$y = -\frac{A}{B}x - \frac{C}{B}.$$

This shows that the locus of the equation

$$Ax + By + C = 0$$

is a straight line with slope equal to $-A/B$, and y-intercept equal to $-C/B$.

If $B = 0$ the equation takes the form

$$Ax + C = 0,$$

which is a straight line parallel to the y-axis. Its slope is infinite and its y-intercept is infinite.

For example,

$$2x - 3y - 6 = 0$$

is the equation of a straight line with slope

$$m = -\frac{2}{-3} = \frac{2}{3},$$

and y-intercept

$$b = -\frac{-6}{-3} = -2.$$

This line is the line 1 in Fig. 15. Its equation may be written

$$y = \tfrac{2}{3}x - 2.$$

Line I is the graph of the function

$$\tfrac{2}{3}x - 2.$$

27. Intercept Form.—Solve the equation

$$\frac{x}{a} + \frac{y}{b} = 1$$

for y, giving

$$y = -\frac{b}{a}x + b.$$

This is a straight line with slope $m = -b/a$, and y-intercept b.

Putting $y = 0$ in the equation and solving for x gives the x-intercept equal to a.

$$\frac{x}{a} + \frac{y}{b} = 1 \tag{3}$$

s called the *intercept form* of the equation of a straight line. The equation of line I (Fig. 15) is at once put into this form by transposition and division by 6, giving

$$\frac{x}{3} - \frac{y}{2} = 1,$$

howing the x-intercept to be 3 and the y-intercept to be -2.

If the x-intercept is 3 and the y-intercept 2, as for line II Fig. 15), the equation of the line is at once written,

$$\frac{x}{3} + \frac{y}{2} = 1.$$

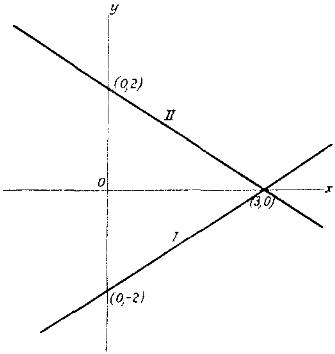


FIG. 15.

Solving for y gives

$$y = -\tfrac{2}{3}x + 2.$$

Therefore,

$$m = -\tfrac{2}{3},$$
$$b = 2.$$

28. Increasing and Decreasing Linear Functions.—Notice that he slope, $-\tfrac{2}{3}$, of line II is negative. This means that the line lopes down to the right. In other words the function

$$-\tfrac{2}{3}x + 2$$

ecreases with increase of x.

For line I the slope $\frac{2}{3}$ is positive, showing that the line slopes up to the right. Therefore, the function

$$\tfrac{2}{3}x - 2$$

increases with increase of x. A positive slope gives an increasing function. A negative slope gives a decreasing function.

29. Straight Line through Two Points.—Suppose the coordinates of two points are given and it is desired to write the equation of the straight line through these points. Let the points be P_1 (1, 2) and P_2 (3, 4) of Fig. 16. Let P, with coordinates x and y, be any point on the required line. The line through P_1 and P

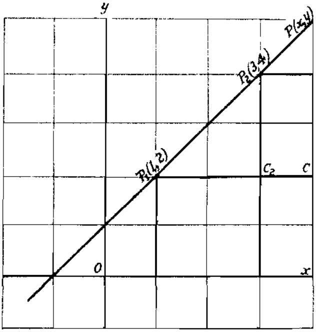


FIG. 16.—Straight line through two points.

must be coincident with the line through P_1 and P_2. Therefore, the slope of P_1P must equal the slope of P_1P_2. By inspection of the figure, and by the rule given on page 59, the slope of P_1P is

$$\frac{CP}{P_1C} = \frac{y - 2}{x - 1},$$

and the slope of P_1P_2 is

$$\frac{C_2P_2}{P_1C_2} = \frac{4 - 2}{3 - 1}.$$

Since these two slopes are equal, we write

$$\frac{y - 2}{x - 1} = \frac{4 - 2}{3 - 1}.$$

Solving for y gives

$$y = x + 1.$$

In general, if P_1 is the point (x_1, y_1) and P_2 is (x_2, y_2), the equation of the line through P_1 and P_2 is

$$\frac{y - y_1}{x - x_1} = \frac{y_2 - y_1}{x_2 - x_1}.$$

This can be readily reduced to any one of the standard forms previously given.

In a later chapter there will be occasion to use the equation of a straight line in getting the straight line of closest fit to a given set of observed data.

30. Circle.—The distance of the point P_1 (3, 4) from the origin in Fig. 17 is $\sqrt{3^2 + 4^2} = 5$. The distance r, of any point $P(x, y)$

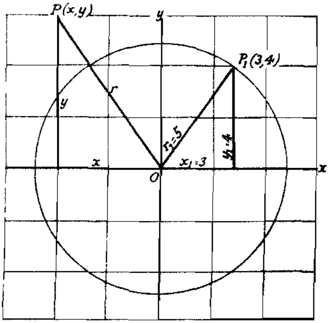


FIG. 17.

from the origin, is $\sqrt{x^2 + y^2}$. If the point P may take only those positions such that its distance from the origin is 5, the equation

$$\sqrt{x^2 + y^2} = 5$$

holds true for all such points and no others. The set of points fulfilling the condition that each is at a distance 5 from the origin forms a circle with radius 5 and center at the origin. So

$$\sqrt{x^2 + y^2} = 5$$

is the equation of a circle with radius 5 and center at the origin. The equation is usually written in the form

$$x^2 + y^2 = 25.$$

If the radius is r, the equation is

$$x^2 + y^2 = r^2.$$

If the center is not at the origin, the equation is easily derived. It is

$$(x - a)^2 + (y - b)^2 = r^2,$$

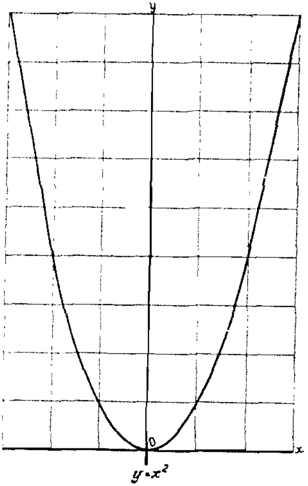


FIG. 18.—Parabola.

for which the center is at the point (a, b). By expanding, the equation becomes

$$x^2 + y^2 - 2ax - 2by + a^2 + b^2 - r^2 = 0.$$

Notice that it has first degree terms in x and y as well as second-degree terms.

31. Parabola.—Examine the equation

$$y = x^2$$

to see what may be discovered in regard to its locus. When $x = 0$, $y = 0$, the locus must pass through the origin. When $x = 1$, $y = 1$, when $x = 2$, $y = 4$, when $x = 3$, $y = 9$, etc. As x increases, y increases more rapidly than does x. In fact, it increases as the square of x. The larger is x, the faster y increases. Negative values of x give the same values of y as do the numerically equal positive values of x. There are no values of x that will give negative values of y. Why?

It is seen that the locus must go through the origin, extending upward both to the right and to the left, getting steeper as it goes up. The right-hand part is symmetrical to the left-hand part with respect to the y-axis. There is no part of the curve below the x-axis. This curve is called a parabola (see Fig. 18).

The locus of the equation

$$y = ax^2$$

will also be a parabola. A point on this curve with a given abscissa will have an ordinate a times the ordinate of that point on the locus of

$$y = x^2$$

that has the same abscissa.

The line of symmetry, in this case the y-axis, is called the axis of the parabola. Where the parabola intersects its axis, in this case the origin, is called the vertex of the parabola.

32. Second-degree Function of x.—If the vertex is not at the origin, the form of the equation changes. It can be shown that if the vertex is at the point (h, k), and with the axis parallel to the y-axis of coordinates, the equation changes from

$$y = ax^2$$

to

$$(y - k) = a(x - h)^2.$$

Expanding, transposing, and solving for y gives

$$y = ax^2 - 2ahx + ah^2 + k.$$

This function of x has a term in x^2, a term in x and constant terms, a second-degree function. The equation may then be written in the form

$$y = ax^2 + bx + c,$$

the general equation of a parabola with its axis parallel to the y-axis.

33. Use of Parabola.—Many sets of observed data which cannot be fitted by a straight line can be quite closely fitted by a

parabola. It may be recalled that the formula giving the distance s through which a body falls under the influence of gravitation in time t is

$$s = \frac{1}{2}gt^2.$$

This function plotted gives a parabola, since it is a second-degree function of t (see Fig. 12). What would be the meaning for negative values of t?

34. Uniform and Non-uniform Scales.—So far, all the plotting of values of variables has been done with a *uniform* scale. By this is meant that for equal differences in the variable there are equal segments of the scale line. On the uniform or arithmetic scale the distance from $x = 1$ to $x = 2$ equals the distance from $x = 3$ to $x = 4$, and so on indefinitely. Many times a *non-uniform* scale is used in order to simplify the plotting of the graph and to facilitate the study of the nature of the function.

The idea may easily be illustrated by means of the parabola equation

$$y = \frac{1}{2}x^2.$$

In this equation put z for x^2 and the equation takes the form

$$y = \frac{1}{2}z.$$

This is linear in z. Tabulate values of z and y for a series of values of x.

TABLE XXIX

$z = x^2, \quad y = \frac{1}{2}z$

x	z	y
0	0	0
1	1	½
2	4	2
3	9	4½
4	16	8

35. Scale of Squares.—In Fig. 19 on the x-axis are located points at distances from the origin equal to the values of z tabulated in Table XXIX. These points are marked with the corresponding values of x shown in the table. Vertical lines have been

drawn at these points for convenience in plotting. The x-scale thus derived is not a uniform scale. The distance from $x = 0$ to $x = 1$ is less than from $x = 1$ to $x = 2$, which, in turn, is less than from $x = 2$ to $x = 3$. If this scale is continued, the scale distances for equal x-differences will continue to increase as x increases. *This may be called a scale of squares.* The actual distance from the origin of any point on the x-axis is the square of the number with which that point is marked.

Now plot, on the x- and y-scales, the pairs of values of x and y as tabulated, paying no attention to the values of z. With these scales available for plotting, it is superfluous to tabulate values of z. The points are marked with small circles. These points appear to be in a straight line. In fact, since the parabolic func-

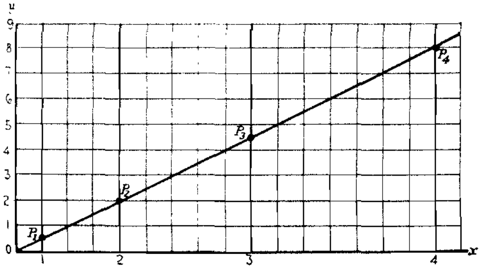


FIG. 19.—Scale of squares.

tion $\frac{1}{2}x^2$ is a square function, these points must fall in a straight line, being plotted to a scale of squares. This line, moreover, is the locus of the equation $y = \frac{1}{2}z$ and must be straight since $\frac{1}{2}z$ is a linear function.

36. Inverse Process.—Suppose the tabulated values of x and y had been the result of observed measurements. Then if the pairs of values, plotted on paper ruled horizontally in a scale of squares and vertically in a uniform scale, determine a set of points in a straight line, it is known that, so far as the observed values are concerned, y is a parabolic function of x. If this line goes through the origin, it is known the equation of the parabola fitting the observed values is of the form

$$y = ax^2.$$

The slope of the line is a. Why? To determine a, take two points on the line and divide the difference of their ordinates by the difference of their abscissas, both measured in the same units of measurement. For instance, the difference of the ordinates of P_2 and P_4 is 6. The difference of their abscissas is 12. The slope of the line is then $6/12 = 1/2 = a$. This fixes the equation of the parabola as

$$y = 1/2x^2.$$

For negative values of x, will the straight line of Fig. 19 continue straight on through the origin or will it change its direction? What is the significance of the result?

37. Establishing a Law.—If the values of x and y in Table XXIX are the result of the observed measurements, there are too few of them to warrant assuming that a law has been established for all similar measurements. If, however, a large number of observations be made, or many sets of observations taken, and always the resultant graph falls very nearly on the line P_1P_4 of Fig. 19, the probability is very great that the law as stated by the equation

$$y = 1/2x^2$$

is established. It would then be quite safe to predict the value of y for any assumed value of x within the limits of the observations.

38. Example.—For example, suppose that a large number of observations of s, the space passed through by a falling body in time t, be tabulated and plotted on uniform scales for both s and t. If the resulting points seem to be nearly on a curve that looks like a parabola, they can then be plotted on a uniform s-scale and a scale of squares for t. If it is found these points fall on a straight line, our hypothesis of the parabola is confirmed. At any rate, by getting the straight line of closest fit to these points, the parabola of closest fit to the original observations may be determined.

39. Logarithmic Scale.—Another very useful non-uniform scale is the *logarithmic* scale.[1] A function which a large class of observed data follows closely is ka^x, giving the equation

$$y = ka^x.$$

In which k and a are constants. Taking logarithms of both sides of the equation gives

$$\log y = x \log a + \log k.$$

[1] A brief discussion of logarithms may be found in Appendix A.

Put

z for the variable $\log y$.
m for the constant $\log a$,
b for the constant $\log k$.

and the logarithmic equation takes the form

$$z = mx + b.$$

This is a linear equation and with uniform scales for x and z will plot as a straight line. If, then, the uniform scale for z be replaced by a logarithmic scale for y, the original equation will plot as a straight line.

40. **Example.**—Take, for example,

$$y = 2^x.$$

Then

$$\log y = x \log 2.$$

Put

$$z = \log y, \log 2 = 0.301.$$

Therefore,

$$z = 0.301\ x,$$

and

$$y = \text{antilog } z.$$

Tabulate values of z and y for a series of values of x.

TABLE XXX

$z = 0.301x$, $y = \text{antilog } z$

x	z	y
0	0	1
1	0.301	2
2	0.602	4
3	0.903	8
4	1.204	16

Values of z were found and tabulated by multiplying the corresponding value of x by 0.301, the logarithm of 2. Each value of y was obtained from a table of logarithms by looking up the number of which the corresponding value of z is the logarithm, that is, by finding antilog z.

Of course, the values of y may be computed directly from the equation

$$y = 2^x$$

by finding powers of 2 corresponding to each value of x.

In Fig. 20 the x-scale and the z-scale are uniform. The equation,

$$z = 0.301\,x$$

being linear, will plot to these scales as a straight line. A non-uniform y-scale is constructed so that the distances from the origin marked with values of y are equal to the logarithms of these

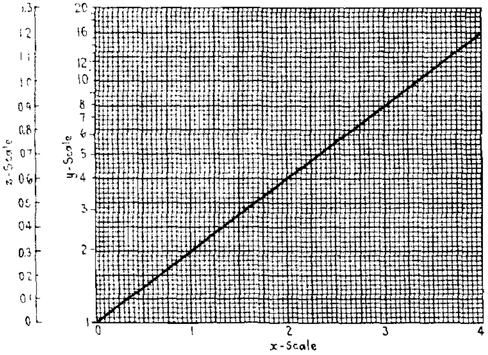


FIG. 20.—Logarithmic scale.

values as measured on the uniform z-scale. This y-scale is a *logarithmic* scale. Suppose that the values of z have not been introduced, but the values of y have been computed directly from the equation

$$y = 2^x,$$

and tabulated with the values of x. If the pairs of values of x and y be plotted on the uniform x-scale and the logarithmic y-scale, the straight line of Fig. 20 is obtained at once.

41. Exponential.—If the values of x and y are both plotted to uniform scales, the exponential curve of Fig. 21 is obtained. For negative values of x, this curve gets closer and closer to the

x-axis but never reaches it. If a set of observed measurements when plotted seems to follow this sort of curve, it may be tested by plotting on a uniform horizontal scale with a logarithmic vertical scale. If the points fall on a straight line, the observed data follow the exponential law.

Paper ruled with uniform scale in one direction and logarithmic scale in the other direction is now easily obtained in the market. It is almost indispensable in studying and comparing rates of

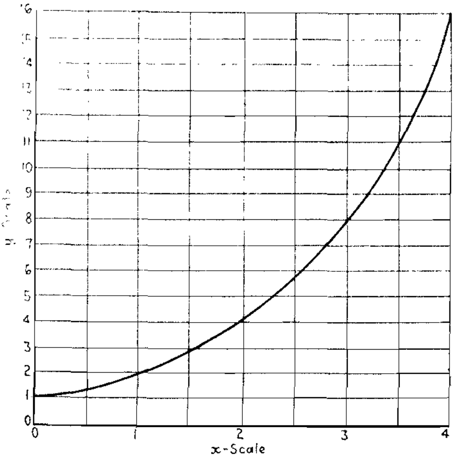


Fig. 21.—Exponential curve.

growth or percentages of change. The subject will be taken up in more detail in a later chapter.

42. Frequency Graphs.—A very important form of graphic representation is that of the frequency distribution. The graph is called a *histogram*. Do not confuse this term with *historigram*. The word "historigram" applies only to historical variables, that is, to those where time is the independent variable.

The form and the construction of the histogram are best shown by an actual example. Take the distribution of leaf lengths as

given in Table XXIII, plotting size of item on the horizontal scale and frequency on the vertical scale.

43. Rectangular Histogram.—The rectangular histogram is quite similar to the vertical-bar diagram. Suitable scales must be chosen so that the graph of the data may be made to fit the paper, be of sufficient size to be read easily, and be of good proportions. The width of the rectangles is equal to the class interval. The left-hand side of the first rectangle is plotted at the lower *boundary* of the first class, bringing sides of the rectangles *at class boundaries throughout the table.* The right-hand side

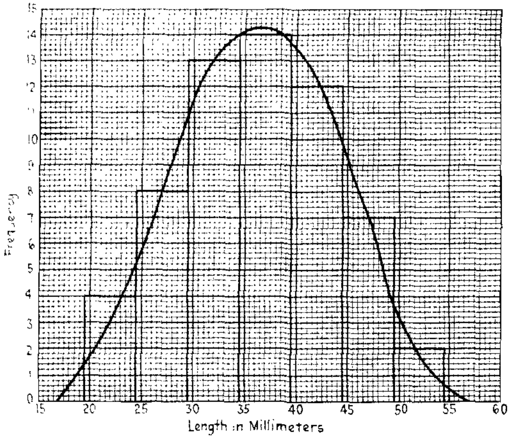


FIG. 22.—Rectangular histogram of sixty leaf lengths.

of the last rectangle must then come at the upper boundary of the last class. The height of each rectangle is made equal to the frequency of the class it represents. The finished rectangular histogram is shown in Fig. 22.

44. What Is Shown.—It can now be seen at a glance of the eye:

1. The manner in which frequencies increase and decrease.
2. The sizes at which the frequencies tend to concentrate.
3. The class of greatest frequency.

4. Whether there is more tendency to long or to short leaves. It is noticeable that the frequencies increase up to about the middle class and then decrease to the end. The maximum frequency occurs in the middle class.

45. Meaning of Area.—If the class interval is taken as the unit of *horizontal measurement and unit frequency as the unit of* vertical measurement, the area of the rectangular histogram up *to the upper boundary of any class is then the total* frequency up to and including that class.

46. Frequency Polygon.—Another *form of* histogram *is* the frequency polygon. This is shown in Fig. 23. Scales are chosen

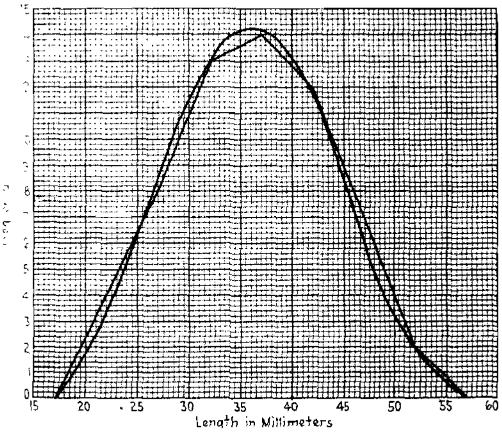


Fig. 23.—Frequency polygon of sixty leaf lengths.

as was done for the rectangular histogram. The values of the frequencies are plotted as points at the midpoint of their respective classes. These points are connected by straight lines from each to the next. In the tabulated sample of 60 leaf lengths, the frequency of class 15–19 is 0. So the polygon is drawn to the base line at 17, the midpoint of this class. Every class smaller than this has a frequency of 0; so the polygon stops at 17 on the base

line. Similarly, for the other end of the table, the polygon stops at 57.

47. Area.—The area of the polygon is exactly the same as that of the rectangles. This may be readily seen if the polygon is drawn by connecting the midpoints of the tops of the rectangles and extending it to the midpoints of the next class above and below those represented by the rectangles. It follows that, if units are chosen as was done for the rectangles, the area of the polygon up to the midpoint of any class equals the total frequency up to and including half of that class.

The fact that the polygon extends beyond the limits of the table suggests that, if the lengths of all the leaves on the tree, from which the sample of 60 leaves was taken, had been measured, a few would have been found shorter than any in the sample and also a few longer. It is likely that this was true.

The frequency polygon is a line or peak-top diagram. It is probably easier to see the tendencies in variation of frequency in the polygon than in the rectangular histogram.

48. Distribution Shown by Sample.—Both the polygon and the rectangles show the variation of frequencies for the 60 leaf lengths actually measured. The presence or absence of irregularities of distribution in the sample is shown. If one is interested in these 60 only, the representation is complete. But the purpose of the investigation would usually be to ascertain the manner in which frequencies are distributed over various sizes for all the leaves on the tree or perhaps for all the leaves of the particular variety used. For this purpose, the sample used must be a *random sample*, so chosen as to be representative of all leaf lengths of the entire set it is desired to study. The entire set which the sample is supposed to represent is called the *universe* or *population* from which the sample is taken. A sample of 60 leaf lengths is of too small a number to be reliable as a random sample.

49. Frequency Curve.—Assume, however, that this sample of 60 is fairly representative for the population from which it was chosen, say all the leaves on the tree, a large number. Then to get a graph for the entire population, a smooth curve is drawn, based on either the rectangular histogram or the frequency polygon. This smooth curve is called a *frequency curve*. In drawing this curve, bear in mind that its purpose is to bring out *general tendencies* of the entire universe represented by the sample. This curve should meet the base line at or near the same

points where the frequency polygon meets it. It goes up to the right, relatively slow at first, and then gets steeper to a certain point where it becomes less steep until it reaches a high point and turns down. It now gets steeper until a point is reached where it flattens out again to meet the base line. This distribution suggests the *symmetric binomial* distribution of Chap. II. If leaves x mm. shorter than the average are equally likely to occur with leaves x mm. longer than the average, and if medium-length leaves are most likely to occur, while the shortest leaves and the longest leaves are few in number, it is justifiable to assume that the symmetric binomial distribution would fit fairly well. The theoretical frequencies could then be computed from the coefficients of the expansion of $(\frac{1}{2} + \frac{1}{2})^6$, and the curve plotted. From the symmetry of the coefficients, the curve would be a symmetric, bell-shaped curve.

Now, it is not known that leaf length should be distributed according to the above-mentioned law. In fact, it seems reasonable that in some varieties of leaves there would be more of a tendency to long leaves or, in others, short leaves, than is shown by this sample. Most frequency distributions are of such a nature that some one of a set of type forms of curves can be found to fit the distribution with a fair degree of closeness. The study of these type curves requires more advanced mathematics.

50. Drawing the Smooth Curve.—It will be necessary, here, to draw the curve so that, exercising the best judgment, it represents the distribution of frequencies in the entire population from which the sample was selected. If it is known that the data are the result of natural phenomena or of pure chance, it is justifiable to smooth out irregularities, assuming that they are accidents of the sample. This matter could be tested out by collecting other samples and examining them to see whether the same irregularities occur at the same size of item. The larger the sample, if properly selected, the more likely it is to be representative of the entire population. If cumulative errors are present, increasing the number of items in the sample is of no use. In the case of economic or sociological data, irregularities are quite likely to occur that it would not be justifiable to smooth out.

In most cases the high point of the curve will occur in the same class as the high point of the polygon or the highest rectangle. If, as in this case, the next lower class has greater frequency than the next higher class, the high point of the curve may be placed

below (to the left of) the middle of the class of greatest frequency. The histogram readily shows the relative weights of these frequencies next to the class of greatest frequency.

Sometimes the curve should go higher than the highest point of the polygon or the highest rectangle, sometimes not so high. From the heights of the third, fourth, and fifth rectangles of Fig. 22, it would seem likely that the sample chosen has too many items in the third class, not enough in the fourth class, and perhaps too many in the fifth class, to be truly representative of all the leaf lengths of the entire tree. If that is so, the curve should lop off, to some extent, the heights for the third and fifth classes, and go above that for the fourth class as shown. In some cases the high point of the frequency polygon will tower above the adjacent points plotted, making it evident that this high frequency is an accident of the sample. If so, the smooth curve should cut off this high peak, bringing out what is likely to be the general tendency.

51. Area of Curve.—The area under the curve should be as nearly as possible the same as that of the rectangles or the polygon. This makes the parts cut off from the rectangles or the polygons equal to the parts filled in outside them. This is similar to the problem of making a railroad grade. The dirt taken from the cuts should even up the fills, and dirt should not be hauled farther than necessary. It should be unnecessary to carry dirt over the top of the hill to the other side. Any number of curves having the correct area could be drawn. Necessarily only one of them is the correct one, giving the true distribution of the universe from which the sample was taken.

52. Histograms from Pure Chance.—Figure 24 shows a frequency polygon of the number of times each possible number of heads occurred in 500 random throws of seven dimes as tabulated in Table XV.

Strictly, only the points plotted are used, since there could not be a fractional number of heads. Number of heads forms a discrete series. The points are connected, however, by straight lines in order to guide the eye in studying the frequency distribution.

The theoretical frequencies, shown in Table XVII, column f', are plotted in Fig. 24 and a smooth curve drawn through the points. This curve is very close to what one would have obtained by smoothing in the same manner as was done, for leaf lengths in

Fig. 23. Comparison of the polygon and the smooth curve shows what adjustments are to be made in the actual observations to get the theoretical results due to the laws of probability.

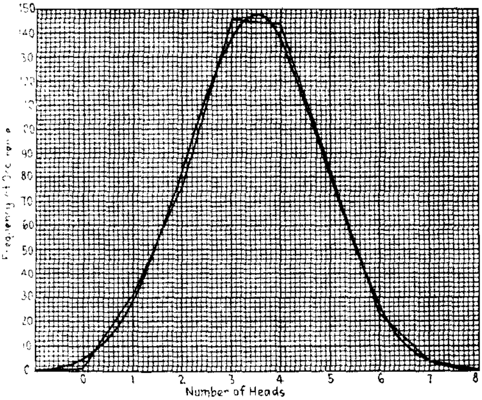


Fig. 24.—Frequency polygon of number of heads in 500 throws of 7 dimes.

53. Symmetric Binomial Histograms.—The theoretical frequencies are the terms in the binomial expansion of

$$(\tfrac{1}{2} + \tfrac{1}{2})^7$$

each multiplied by the total number of throws (see Chap. II, pp. 31 to 33). If there were eight dimes, the theoretical frequencies would be the terms of

$$(\tfrac{1}{2} + \tfrac{1}{2})^8$$

each multiplied by the total number of throws. In general, if there are n dimes, the theoretical frequencies are the terms of the expansion of

$$(\tfrac{1}{2} + \tfrac{1}{2})^n$$

each multiplied by the total number of throws. In every case the sum of the frequencies, that is, the sum of the ordinates of the points plotted, equals the total number of throws of the coins. This sum, then, may be kept constant. If the same horizontal scale is kept as n increases, the width of the polygon, its spread, increases proportionately. It is conceivable that some function

of this spread may be devised such that it may be used as a unit of measurement on the horizontal scale each way from the maximum ordinate. Notice that the maximum ordinate stands at the midpoint of the base line, when $p = q = \frac{1}{2}$. Several such functions of the spread have been used. The one most used is designated by the Greek letter small sigma, σ.

$$\sigma = \sqrt{npq}.$$

A fuller discussion of the derivation and meaning of this function will be given in the chapter on dispersion. Each frequency plotted corresponds to a term of the expansion of $(\frac{1}{2} + \frac{1}{2})^n$. Let x equal the number of terms from the middle of the series, negative if to the left and positive if to the right of the middle.

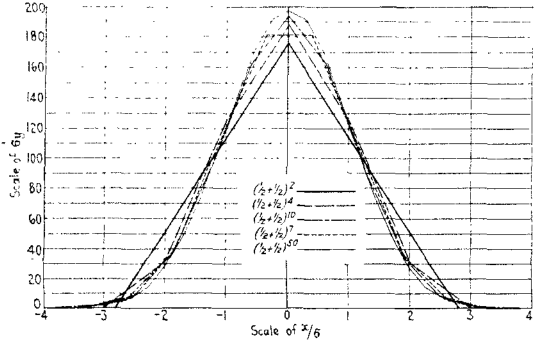


FIG. 25.—Symmetric binomial frequency graphs.

Using the measure σ, the abscissas of points plotted will be x/σ. Then, if each ordinate is multipled by σ, the areas under successive polygons, as n increases, will be the same.

Beginning with $(\frac{1}{2} + \frac{1}{2})^2$, $(n = 2)$, and increasing n, several polygons are constructed in Fig. 25 on the basis of 500 throws. The computations for coordinates of the points are tabulated as follows:

$$n = 2, \quad \sigma = \sqrt{npq} = \frac{\sqrt{2}}{2} = 0.707, \quad \frac{1}{\sigma} = 1.414.$$

$$\left(\frac{1}{2} + \frac{1}{2}\right)^2 = \frac{1}{2^2}(1 + 2 + 1) = 0.25 + 0.50 + 0.25.$$

$n = 2$

x	$\frac{x}{\sigma}$	Per cent y is of 500	y	σy
−2	−2.8	0	0	0
−1	−1.4	25	125	88
0	0	50	250	176
1	1.4	25	125	88
2	2.8	0	0	0
		100	500	

$$n = 4, \qquad \sigma = \sqrt{4 \times \frac{1}{2} \times \frac{1}{2}} = 1, \quad \frac{1}{\sigma} = 1$$

$$\left(\frac{1}{2} + \frac{1}{2}\right)^4 = \frac{1}{2^4}(1 + 4 + 6 + 4 + 1) = 0.0625 + 0.2500 + 0.3750 + 0.2500 + 0.0625.$$

$n = 4$

x	$\frac{x}{\sigma}$	Per cent y is of 500	y	σy
−3	−3	0.00	0	0
−2	−2	6.25	31	31
−1	−1	25.00	125	125
0	0	37.50	188	188
1	1	25.00	125	125
2	2	6.25	31	31
3	3	0.00	0	0
		100.00	500	

$$n = 10, \quad \sigma = \sqrt{10 \times \frac{1}{2} \times \frac{1}{2}} = \frac{\sqrt{10}}{2} = 1.581, \quad \frac{1}{\sigma} = 0.632.$$

$$\left(\frac{1}{2} + \frac{1}{2}\right)^{10} = \frac{1}{2^{10}}(1 + 10 + 45 + 120 + 210 + 252 + 210 + 120 + 45 + 10 + 1)$$

$= 0.000977 + 0.009766 + 0.043945 + 0.117187 + 0.205078 + 0.246094 + 0.205078 + 0.117187 + 0.043945 + 0.009766 + 0.000977.$

$n = 10$

x	$\frac{x}{\sigma}$	Per cent y is of 500	y	σy
−6	−3.795	0.00	0	0
−5	−3.162	0.10	0.5	1
−4	−2.530	0.98	5	8
−3	−1.897	4.39	22	35
−2	−1.265	11.72	59	93
−1	−0.632	20.51	102	161
0	0	24.61	123	194
1	0.632	20.51	102	161
2	1.265	11.72	59	93
3	1.897	4.39	22	35
4	2.530	0.98	5	8
5	3.162	0.10	0.5	1
6	3.795	0.00	0	0
		100.02	500	

$n = 50, \quad \sigma = \sqrt{50 \times \frac{1}{2} \times \frac{1}{2}} = \frac{\sqrt{50}}{2} = 3.535, \quad \frac{1}{\sigma} = \frac{\sqrt{2}}{5} = 0.2828$

$\left(\frac{1}{2} + \frac{1}{2}\right)^{50} = \frac{1}{2^{50}}(1 + 50 + 1{,}225 + 19{,}600 + 230{,}300 + 2{,}118{,}760 + 15{,}890{,}700$ $+ 99{,}884{,}400 + 536{,}878{,}650 + 2{,}505{,}433{,}700 + 10{,}272{,}278{,}170$ $+ 37{,}353{,}738{,}800 + 121{,}399{,}651{,}100 + 354{,}860{,}518{,}600 + 937{,}845{,}656{,}300 + 2{,}250{,}829{,}575{,}120 + 4{,}923{,}689{,}695{,}575 + 9{,}847{,}379{,}391{,}150 + 18{,}053{,}528{,}883{,}775 + 30{,}405{,}943{,}383{,}200$ $+ 47{,}129{,}212{,}243{,}960 + 67{,}327{,}446{,}062{,}800 + 88{,}749{,}815{,}264{,}600 + 108{,}043{,}253{,}365{,}600 + 121{,}548{,}660{,}036{,}300 + 126{,}410{,}606{,}437{,}752$ + each of the numbers preceding this last one, but coming in opposite order).

The last term written above is ${}_{50}C_{25}$, the middle term of the expansion.

$n = 50$

x	$\frac{x}{\sigma}$	Per cent y is of 500	y	σy
−14	−3.96	0.003	Less than 0.1	
−13	−3.68	0.01	Less than 0.1	
−12	−3.39	0.03	0.15	0.5
−11	−3.11	0.08	0.4	1.4
−10	−2.83	0.20	1	4
−9	−2.55	0.44	2	7
−8	−2.26	0.88	4	14
−7	−1.98	1.60	8	28
−6	−1.70	2.70	13.5	48
−5	−1.41	4.18	21	74
−4	−1.13	6.00	30	106
−3	−0.85	7.89	39	138
−2	−0.57	9.64	48	170
−1	−0.28	10.80	54	191
0	0.00	11.25	56	197

Positive values of x repeat all values tabulated in reverse order. The table completed would give 53 values of x from $x = -26$ to $x = 26$. For $x = 26$ and $x = -26$, $y = 0$. For values of x between $x = 12$ and $x = 26$, the values of y are too small to plot unless a very large scale be used.

For $(\frac{1}{2} + \frac{1}{2})^7$ the values of x/σ and σy are as follows computed from Table XVII, p. 33, Chap. II.

$$n = 7, \quad \sigma = \sqrt{7 \times \frac{1}{2} \times \frac{1}{2}} = \frac{\sqrt{7}}{2} = 1.323, \quad \frac{1}{\sigma} = 0.756.$$

$n = 7$

x	$\frac{x}{\sigma}$	σy
$-\frac{9}{2}$	−3.40	0
$-\frac{7}{2}$	−2.64	5
$-\frac{5}{2}$	−1.89	36
$-\frac{3}{2}$	−1.13	108
$-\frac{1}{2}$	−0.38	181
$\frac{1}{2}$	0.38	181
$\frac{3}{2}$	1.13	108
$\frac{5}{2}$	1.89	36
$\frac{7}{2}$	2.64	5
$\frac{9}{2}$	3.40	0

54. Normal Probability Curve.—From the above computation for $(\frac{1}{2} + \frac{1}{2})^{50}$, it is seen that n does not need to be very large to make the labor of computation very great. From the polygons plotted in Fig. 25, it is seen that as n increases from 2 to 50 the sides of the polygons become shorter and seem to be approaching closer and closer to a definite curve. It is true that the polygon determined by

$$(\tfrac{1}{2} + \tfrac{1}{2})^n$$

does approach a definite curve as n increases without limit. This is known as the *normal probability* curve and its equation has been found to be[1]

$$y = y_0 e^{-\frac{x^2}{2\sigma^2}}.$$

The y-axis of coordinates passes through the high point of the curve; so the origin of coordinates is at the same place as in Fig. 25. The ordinate at $x = 0$ is y_0, the maximum ordinate of the curve; e is the base of natural logarithms; σ is the constant used above. With $n = 50$ the polygon is so near the curve that it was deemed best, in order to avoid confusion, not to draw the curve in this figure. It evidently will be slightly above the polygon for $n = 50$, in the middle portion from about $x/\sigma = -1$ to $x/\sigma = +1$. In fact, from the equation of this curve, it can be proved by calculus that the direction of curvature changes at the points where $x/\sigma = -1$ and $x/\sigma = +1$. To the left of -1 and to the right of $+1$ the curve will be slightly below the polygon for a considerable distance. The polygon for $n = 50$ does not quite reach the horizontal axis until $x = -26$ and $x = +26$, where $y = 0$. The curve never quite reaches the horizontal axis but gets closer and closer to it as x/σ increases. In tables for statisticians[2] the values of the ordinate y are tabulated for values of x/σ. At $x/\sigma = 4$, y is only 0.013 per cent of the total frequency.

The normal probability curve is also known as the normal frequency curve, the normal curve of error, the Gaussian curve, the Gauss-Laplace curve. Since the distributions of many frequencies follow this approximately, its properties are of great

[1] G. Udney Yule, "Theory of Statistics," p. 301, does not use calculus. Truman L. Kelley, "Statistical Method," p. 94, gives a simple derivation using calculus.

[2] "Tables for Statisticians and Biometricians," Karl Pearson. "Tables of Applied Mathematics," James W. Glover.

importance in the statistical treatment of observations. It will be used later in this book.

55. Rectangular Histogram of Weights of Freshmen.—Now construct a rectangular histogram of the weights of freshmen as shown in the frequency table, Table XII (Chap. II). Choose proper scales to make the figure well proportioned. Remember to place scale marks on the heavy lines of the ruling. Round numbers are chosen such that subdivisions of the scale are easily read. Figure 26 shows the result. It will be noticed that the

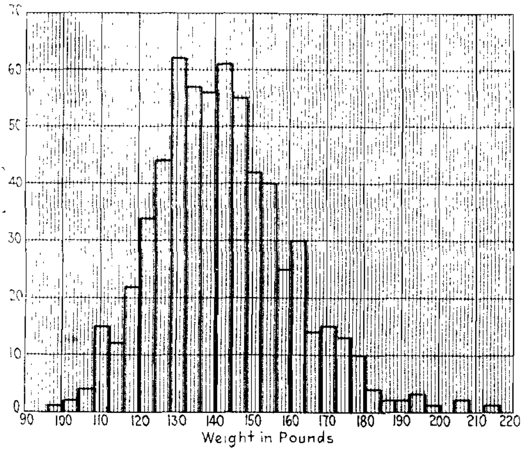


FIG. 26.—Rectangular histogram of weights of freshman men. (Small class interval.)

histogram tails off to low frequencies in the heavy-weights. If this *tail*, as it is called, were cut off, the remaining part of the histogram would be fairly symmetrical and not far from a symmetric binomial distribution. There are, however, irregularities. For example, if the top of the fourth rectangle were cut off and added to the fifth, the distribution would be much smoother. It is likely that this irregularity is merely an irregularity of the sample which another sample would not show. A similar situation is seen in the neighborhood of weights from 160 to 170 lb. There

seem to be two high points. Again it is a question whether another sample would show the same thing. Perhaps the two high points should be cut off and distributed to the two lower ones between them. It is a matter which must be decided if a smooth curve is to be drawn.

56. Frequency Polygon of Weights of Freshmen.—Using the same data, Fig. 27 shows the frequency polygon. The irregularities above mentioned now stand out more conspicuously than in the rectangular histogram. If now a smooth curve were made, it would almost surely be drawn so as to smooth out all the irregu-

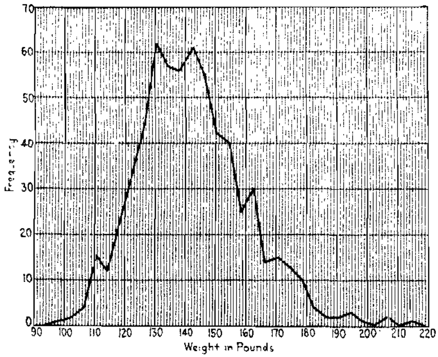


FIG. 27.—Frequency polygon of weights of freshman men. (Small class interval.)

larities unless there is question in regard to whether there should be two high points or not. In connection with the frequency tables, it was noted that increasing the class interval wipes out irregularities of the sample and brings out general tendencies.

57. Effect of Wider Class Interval.—Proceed next to graph Table XIII. Figure 28 shows the rectangular histogram, and Fig. 29 the frequency polygon. The smoothing out of the irregularities due to increasing the class interval is very marked. It is now a very simple matter to draw a free-hand smooth curve fitting the distribution quite well. The characteristic bell-

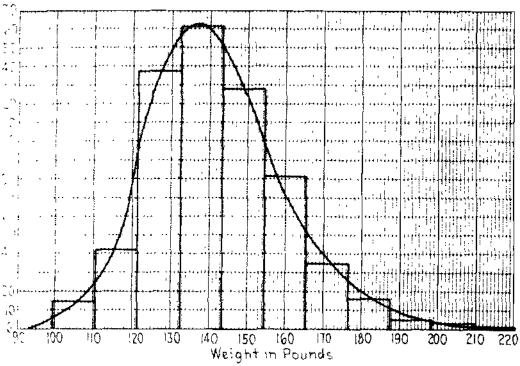


FIG. 28.—Rectangular histogram of weights of freshman men. (Large class interval.)

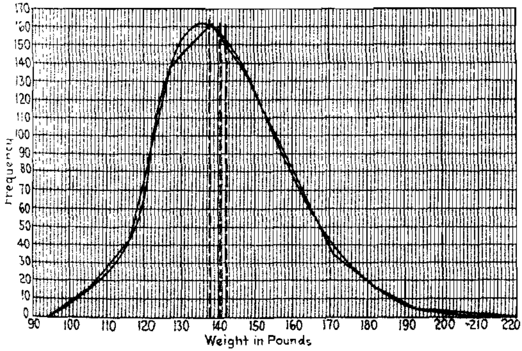


FIG. 29.—Frequency polygon of weights of freshman men. (Large class interval.)

shaped curve is obtained though not quite symmetrical. The tail to the right is considerably longer than to the left.

58. Frequency Graphs of Freshman Heights.—The heights of freshmen as given in Table XIV will now be graphed. The rectangular histogram is shown in Fig. 30 and the frequency polygon in Fig. 31. Slight tendencies to hump up are shown in the polygon at heights 63.5 and 72.5 which are hardly noticeable

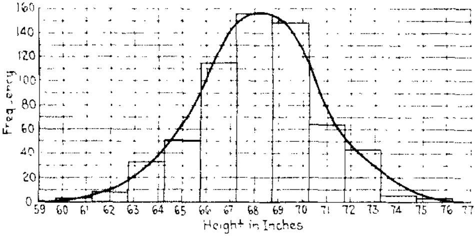


FIG. 30.—Rectangular histogram of heights of freshman men.

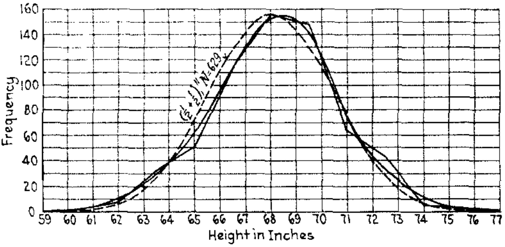


FIG. 31.—Frequency polygon of heights of freshman men.

in the rectangles. It is reasonable to assume that the polygon is too high at height of 69.5, too low at 71, too high at 72.5, and too low again at 74. These and similar considerations on the other half of the polygon assist in smoothing.

59. Comparison with Symmetric Binomial Curve.—If 11 coins were tossed 629 times, the symmetric binomial frequency curve that would be expected by virtue of the laws of probability,

showing the frequency of heads, is shown by the dotted curve in Fig. 31. This originates, of course, from the terms of the expansion of

$$(\tfrac{1}{2} + \tfrac{1}{2})^{11}.$$

The closeness of this curve to the free-hand, smooth curve already drawn is very noticeable. The heights of these 629 freshmen then gives an excellent illustration of the following of a symmetric point binomial frequency distribution by a variable which is the result of natural phenomena.

60. **Ogive.**—Another form of frequency graph called the *ogive* is of considerable importance for some purposes. It is simply a

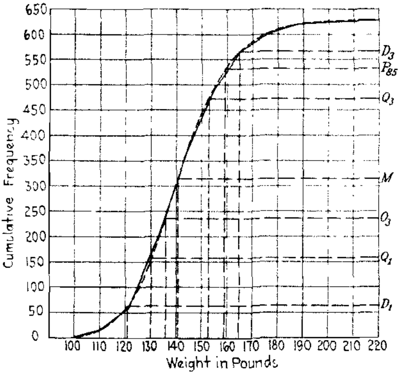


FIG. 32.—Ogive of weights of freshman men.

graph of *cumulative* frequencies. In the frequency polygon the frequencies were plotted at the midpoints of the classes. This may be done on the assumption that the midpoint of a class is the average size of the items occurring in that class. This seems to be the fairest assumption to make if the actual distribution within the class is unknown. In the case of the ogive, the cumulative frequencies are plotted at the *upper boundaries* of the respective classes. This is done because the cumulative frequency at any class is the *total frequency* up to and including that class. In other words, it is the total frequency up to the upper boundary

of the class. The graph starts on the base line at the lower boundary of the first class.

Figure 32 is the ogive of freshman weights plotted from the cumulative frequency column of Table XIII. Figure 33 is that for freshman heights. The angular ogive never goes down to the right, since there are no negative frequencies. It would remain horizontal at classes with zero frequency. Otherwise it is always going up to the right.

Difference of ordinates between any two points gives the increase in number of items between the sizes of item measured by the abscissas of the two points. So the steeper the graph the

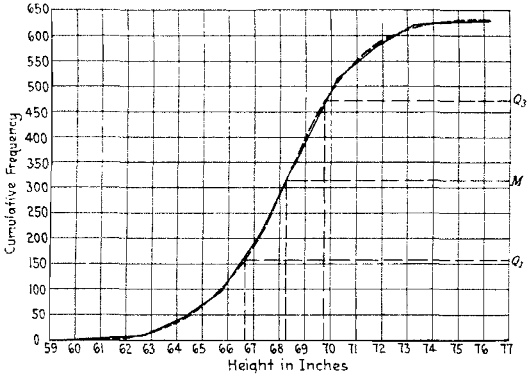


FIG. 33.—Ogive of heights of freshman men.

more rapid the increase of total frequency. With a symmetric bell-shaped distribution, the ogive goes up slowly at first, most rapidly at the middle, and slowly again at the end, being symmetrical with respect to the middle point. This is very nearly true in Fig. 33. In Fig. 32 the flat part at the upper end is much longer than the flat part at the lower end. This corresponds to the long tail to the right in the frequency polygon (Fig. 29).

61. Comparison of Ogive with Histogram.—The ogive is more easily smoothed than the frequency polygon. The curve does not usually have to depart very far from the points plotted from the cumulative frequency column.

In the frequency polygon, the area up to the ordinate of any point (the class interval being the unit of horizontal measure) equals the total frequency up to that point. The area under the smooth histogram up to any ordinate, proper units of measure being used, gives the total frequency up to that ordinate. In the ogive the ordinate at any point gives the total frequency up to that point. At any given value of size of item the ordinate to the ogive equals the area under the smooth histogram up to the same point. Such a curve as the ogive is called the integral curve of the smooth histogram.

Certain division points are easily determined graphically from the ogive. This will be taken up in a later chapter.

62. Statistical Maps.—A form of graph used a great deal is the statistical map. Outline maps are printed for this purpose or such a map may be made. The use of the statistical map is to show geographic distribution. The data may be shown by various colors, by various forms of shading or cross-hatching, or by dots. A wide variety of data may be shown in a variety of ways. Only one form will be discussed here, the *dot map*. Suppose it is desired to show graphically the number of automobiles in the states of the United States. Of course, the data are first tabulated by states. By examining the data a scale may be chosen. The scale is one dot for a certain number of automobiles, perhaps one dot for every 10,000. Then an outline map of the United States, showing outlines of the states, is taken and in each state is placed a number of dots equal to the number of 10,000 automobiles in that state. If the dots are arranged at the same distance apart everywhere, the numbers of automobiles in any two different states are readily compared by the size of the groups of dots in those states. If the dots are scattered uniformly all over the area of each state, then the density of automobiles in different states is readily compared. The farther apart the dots, in this case, the fewer automobiles per square mile.

This example gives some idea of what may be done with the statistical map. The subject is very well treated in books on graphic charts.

63. General Rules.—A few general rules may be given for making charts of various kinds.

1. Choose the scales carefully. They should be chosen so as to cover all the data and make a well-proportioned figure on the paper used.

2. *Scales should be such as to bring out the fluctuations of the function without unduly magnifying them.*

3. The scales should be large enough to enable one to read the graph to the required degree of accuracy.

4. *The scale lines should be ruled as heavy lines.*

5. The scale marks should be placed on the heavy lines of the ruled paper.

6. The scale marks should be integers or round numbers so chosen that the subdivisions of the ruled paper are easily read.

7. In comparing two or more variables, the units of measure should be such as to bring out the fluctuations of each. For example, if the number of pounds of flour and the number of pounds of pepper used by the average family are plotted on such a scale as to bring out the fluctuations in flour used, the fluctuations in pepper will scarcely show. If ounces of pepper and hundredweights of flour are plotted, it is much more likely that the fluctuations in both will be brought out so as to be compared.

8. Neatness and accuracy are prime requisites.

Exercises

1. On millimeter paper plot a horizontal bar chart of the following data of passenger automobiles exported to different countries in 1921.

LEADING FOREIGN MARKETS FOR AMERICAN PASSENGER CARS[1]

Country	Number	
	1921	1922
Australia	3,020	11,236
Canada	5,243	10,214
Mexico	6,750	7,279
Belgium	533	4,785
United Kingdom	888	4,315
Sweden	920	3,063
Argentina	613	2,497
British South Africa	596	2,327
Spain	421	2,117
Other countries	11,966	18,957
Total	30,950	66,790

[1] From "Commerce Yearbook" for 1922.

2. Plot the same data as a vertical-bar chart.

3. Plot a horizontal-bar chart showing the data for both 1921 and 1922. Plot the data for 1921 in black and those for 1922 in red. Place the bars for the two years for each country close together, leaving a wide space between the two bars for one country and the two bars for the next country.

By a glance at the chart tell which country gave the largest per cent of increase for 1922 over 1921.

By a glance of the eye at the chart tell whether exports to any country were less in 1922 than 1921.

Which country took the smallest number of cars in 1921? That country took about how many times as many cars in 1922?

Which country took the smallest number of cars in 1922? That was about how many times as many as that country took in 1921?

4. Make a vertical-bar chart showing the number of passenger cars exported from the United States for each year from 1913 to 1922 inclusive.

UNITED STATES EXPORTS OF PASSENGER CARS[1]

Year	Number
1913	25,880
1914	22,335
1915	41,864
1916	61,922
1917	65,756
1918	36,936
1919	67,145
1920	142,508
1921	30,950
1922	66,790

[1] From "Commerce Yearbook" for 1922.

In this chart does the increase of number of automobiles exported in 1922 over that of 1921 look as impressive as it did in the chart of Ex. 3? Why?

5. Make a horizontal-bar chart of the following freight data. Make one bar for each year, subdivided so as to show the amount of freight originating in each of the three districts. Make the snbdivisions for each district different either by three colors or three forms of cross-hatching. Make a legend in a convenient place indicating the meaning of each color or cross-hatching.

Complete the table of data by forming a column of totals for the United States.

REVENUE FREIGHT ORIGINATING ON CLASS I ROADS, BY DISTRICTS[1]
(In short tons)

Calendar year	Districts		
	Eastern	Southern	Western
1916	588,934,108	196,392,311	416,673,648
1917	617,844,385	211,474,636	434,696,704
1918	614,703,873	216,081,819	432,558,301
1919	523,810,514	194,564,103	377,736,654
1920	606,786,167	224,127,173	424,507,651
1921	449,674,156	177,258,034	313,250,370
1922	451,585,046	208,582,180	362,942,352

[1] From "Commerce Yearbook" for 1922.

If a scale is used such that only three-figure accuracy can be plotted, plot the data to the nearest million tons.

For which district is it easiest to compare the changes from year to year? For which district is this comparison most difficult? Why is this?

6. Make a table showing the percentage each entry of the table in Ex. 5 is of the total for the corresponding year.

Make a 100 per cent bar chart of the results, one bar for each year.

For which year did the eastern division originate the smallest percentage of freight? Was its total tonnage that year also less than for any other year?

For which year did the southern division originate the greatest percentage of freight? Was its total tonnage that year also greater than for any other year?

7. Make a pie diagram of car loadings for each quarter of 1922. Make the areas of the circles proportional to the total number of car loadings for each quarter.

NUMBER OF CAR LOADINGS, CLASS I ROADS, BY PRINCIPAL COMMODITIES, FOR 1922[1]

Quarter	Grain	Live stock	Coal	Forest products	Merchandise
First	606,208	379,587	2,379,387	650,817	2,848,223
Second	499,622	372,899	1,064,595	779,355	3,155,667
Third	685,797	396,315	1,494,101	735,605	2,985,878
Fourth	675,731	489,122	2,510,258	773,269	2,888,044

[1] From "Commerce Yearbook" for 1922.

Is there much apparent difference in the sizes of the circles? The total for the third quarter was about one-fourth bigger than that for the second quarter. Do the corresponding circles seem to have about that proportional size?

8. Reduce the number of car loadings of each commodity to percentage of the total number of car loadings for the corresponding quarter. Make a 100 per cent pie diagram for each quarter. These circles should all have the same area.

The number of car loadings of forest products for the second and fourth quarters were about the same. Do the diagrams reveal for which of these two quarters the number of car loadings of forest products was the greater percentage of the total car loadings for the corresponding quarter?

9. Make a line chart of the revenue freight data shown in Ex. 5. Use a different type of line for each district. Make a legend, in a convenient place, showing what each line represents, or label each line in the chart to show which district it represents.

Do the increases and decreases from year to year seem to run about the same for all districts?

Is there any time during the period of seven years when the number of tons for the southern district is greater than that for each of the other districts?

10. Plot a line chart of the following population of the state of Washington.

POPULATION OF WASHINGTON

Year	Population
1860	11,594
1870	23,955
1880	75,116
1890	357,232
1900	518,103
1910	1,141,990
1920	1,356,621

Connect each plotted point to the next by a straight line. After this is done, draw a smooth curve through the plotted points.

Does the graph show whether the increases from decade to decade are themselves increasing? How?

If the population increased the same number each year from 1890 to 1900 how can the population in 1897 be determined graphically?

11. Choose a pair of coordinate axes and plot the following functions:

(*a*) $y = \frac{3x}{2}$, (*b*) $y = \frac{3x}{2} + 2$, (*c*) $y = \frac{3x}{2} - 2$, (*d*) $y = -\frac{3x}{2}$,

(*e*) $y = -\frac{3x}{2} + 2$, (*f*) $y = -\frac{3x}{2} - 2$, (*g*) $y = \frac{2x}{3}$, (*h*) $y = \frac{2}{3}x + 2$.

What are the relative positions of the graphs of (*a*), (*b*), and (*c*)? How could this situation have been seen without plotting the lines?

What seems to be the position of lines (*d*) (*e*) and (*f*) with respect to (*g*)?

12. Where do the loci of $x/3 + y/4 = 1$ and $x/3 - y/4 = 1$ cut the coordinate axes?

13. Find the equation of the straight line through the points (1, 2) and (2, −1). Also find the equation of the straight line parallel to this line and passing through the point (2, 3). Where does this latter line cut the y-axis?

14. How may the locus of $x^2 + y^2 = 16$ be drawn without plotting points?

15. Plot the locus of $y = x^2/2$. What is this curve called? Where is its vertex?

16. Plot the parabola $y = -x^2/2 - 2x + 2$. Where is its vertex? Which way does the curve run from the vertex?

17. Make a non-uniform x-scale so that $y = x^2/2$ shall plot as a straight line.

18. Procure a sheet of semilogarithmic paper and plot on it the population of the state of Washington as given in Ex. 10.

Connect each plotted point to the next by a straight line.

Did the population increase at a more rapid rate during the first decade or during the last decade of the record? What shows this on the graph?

For how long after 1860 did the population increase at an increasing rate? How is this shown?

During which decade was the rate of increase least? What shows this?

By graphic interpolation determine the population in 1895. How would this result be computed from the tabulated data?

19. Make a *rectangular histogram of mean monthly temperatures* at Seattle, using the frequency table of Ex. 2, Chap. II. Superimpose a frequency polygon on the rectangular histogram. Draw a smooth *frequency curve*.

20. Make a rectangular histogram of monthly precipitation at Seattle, using the frequency table of Ex. 3, Chap. II.

Construct a frequency polygon, separately, for the same data. Draw a smooth curve for each figure without comparing either with the other. After the two curves are finished, compare the results.

21. Construct a rectangular histogram of frequencies for all possible number of heads as shown in Table XV.

22. Construct an ogive for the data in each of the three preceding exercises.

23. Make a frequency polygon of the temperature data as shown in the frequency table of Ex. 6, Chap. II. What characteristic difference is there between this graph and the polygon of Ex. 19? What is the reason for this difference?

24. Construct a frequency polygon from the table of Ex. 10, Chap. II. On the same set of axes plot a polygon of the theoretical frequencies of Ex. 11, Chap. II. If 10,000 tosses of the 10 coins had been made and the resulting actual frequencies and theoretical frequencies graphed as above, would the two polygons be likely to be more nearly alike than the two just plotted?

25. Make a horizontal-bar chart of the following data of imports:

United States Imports of Merchandise (Specie Values)

Fiscal year	Free	Dutiable
1913	$987,524,162	$ 825,484,072
1914	1,127,502,699	766,422,958
1915	1,033,526,675	640,643,065
1916	1,492,647,350	705,236,160
1917	1,848,840,520	810,514,645
1918	2,118,599,372	827,056,031
1919	2,230,222,808	865,497,260
1920	3,405,233,003	1,833,119,111
1921	2,137,440,504	1,517,018,842
1922	1,598,888,618	1,009,190,390

26. Make a vertical-bar chart of the data of Ex. 25.

27. Make a line chart of the data of Ex. 25.

28. Compute and tabulate for each year the per cent free imports and the per cent dutiable imports are of the total imports in Ex. 25. Make a per cent chart of the results.

29. Do the two classes of imports increase and decrease together? Which of the charts in Exs. 25 to 28 best shows this? Does the per cent of free imports remain fairly constant or not? What shows this graphically?

CHAPTER IV

AVERAGES

Statistics has been called the science of averages. An *average* is a type form of a group of items used to represent the entire group. In a sense it is a representative sample consisting of one value only. It is not necessary, however, that it actually have the value of one of the items of the group it represents. It would seem that its size should be somewhere about midway in the entire group of sizes.

1. General Definition.—In general, some items of the group are larger than the average and some are smaller. The average will satisfy the condition of being such a function of the entire group of values that if all of the group happen to be equal to each other, then the average equals each one of the group.[1]

2. For Comparison.—Averages make it possible to compare different groups. This was suggested in Chap. I in comparing heights of trees in different forests. Sometimes comparison of averages might be meaningless. This was seen in Chap. I, where it was desired to compare the amount of lumber in two forests by comparing their average tree heights.

3. Arithmetic Mean.—When asked to find the average of a set of numbers, the kind of average that probably comes to mind is the arithmetic mean. The *arithmetic mean* of a set of values, $X_1, X_2, X_3, \ldots X_N$, is their sum divided by the number of items in the set.

$$a = \frac{X_1 + X_2 + X_3 + \ldots + X_N}{N} = \frac{\Sigma X}{N}.$$

In this formula X_1 is the size of the first item in the set, X_2 the size of the second item, and so on to X_N the size of the Nth item. N is the number of items. ΣX is the sum of all such items as X_1, X_2, and so on. This is read "sum of such terms as X" or "sigma X." The arithmetic mean of the N items is a.

[1] Prof. Irving Fisher, "The Making of Index Numbers."

The arithmetic mean satisfies the general definition of average given on page 96. For if $X_1 = X_2 = X_3 = \ . \ . \ . = X_N = k$, the arithmetic mean,

$$a = \frac{\Sigma X}{N} = \frac{Nk}{N} = k.$$

That is, a equals each one of the set.

4. Important Property of a.—If the totality of all values of the items ΣX is known and the number of items N, it is not necessary to know the individual items to get the arithmetic mean.

$$a = \frac{\Sigma X}{N}.$$

Or, if a and N are known, the totality of all values is determined.

$$\Sigma X = aN.$$

If a man travels 30 miles in 10 hr., his average speed is $\frac{30}{10} = 3$ miles per hour. To find the average it is not necessary to know his speed at any one time. If his speed was less than 3 miles per hour at one time, it had to be greater than 3 miles per hour at some other time. He may even have stopped to rest part of the time, when his speed was nil. Or he may have gone backwards part of the time, when his speed was negative.

5. Deviations from a.—A valuable property of the arithmetic mean is that the *algebraic sum of the deviations of all the items from the arithmetic mean is zero.*

By *deviation* of an item from the arithmetic mean is meant the remainder obtained by subtracting the arithmetic mean from that item. Letting d stand for deviation:

$$d_1 = X_1 - a, \ d_2 = X_2 - a, \text{ etc.}$$

To prove the sum of the deviations equal to zero, write them down and add.

$$\begin{aligned} d_1 &= X_1 - a, \\ d_2 &= X_2 - a, \\ d_3 &= X_3 - a, \\ &\ \ \vdots \\ d_N &= X_N - a. \end{aligned}$$

Adding, $\quad \Sigma d = \Sigma X - Na.$

Since

$$\Sigma X = Na,$$
$$\Sigma d = Na - Na = 0,$$

which was to be proved.

6. Deviations from Any Arbitrary Value.—*The average deviation of all the items from any arbitrary number added to that arbitrary number equals the arithmetic mean of the items.*

Letting the arbitrary number be X', the above theorem stated in symbols is

$$a = X' + \frac{\Sigma(X - X')}{N}.$$

The proof is as follows:

$$X' + \frac{\Sigma(X - X')}{N} = X' + \frac{\Sigma X}{N} - \frac{NX'}{N}$$
$$= X' + \frac{\Sigma X}{N} - X'$$
$$= \frac{\Sigma X}{N} = a.$$

Or, more in detail, $\Sigma(X - X')$, the sum of the deviations from X', is obtained by direct addition

$$X_1 - X'$$
$$X_2 - X'$$
$$X_3 - X'$$
$$\cdot$$
$$\cdot$$
$$\cdot$$
$$\cdot$$
$$X_N - X'$$

Adding, $\Sigma X - NX'$.

Therefore,

$$\Sigma(X - X') = \Sigma X - NX',$$

and the above proof follows at once.

7. Short-cut.—This property of the arithmetic mean makes it possible to take deviations from a convenient number and from these compute the mean. This process is known as the *short-cut method* of obtaining arithmetic mean. This may be illustrated

by an example. Find the arithmetic mean of the numbers in the following column:

86,745
86,749
86,758
86,761
86,7
5)433,7
86,755 = a.

Or, letting X' equal the middle number of the set,

X	$d_{X'}$
86,745	− 13
86,749	− 9
86,758	0
86,761	3
86,762	4
$N = 5$	$\Sigma d_{X'} = -15$

$$a = X' + \frac{\Sigma d_{X'}}{N} = 86{,}758 + \frac{-15}{5}$$
$$= 86{,}755.$$

This saved the bother of adding the set of five-figure numbers. Someone may say that since the first three figures are the same for all these numbers, it is only necessary to average the numbers given by the last two figures, and annex the result to the first three figures. This is only the short-cut method, using 86,700 for X'.

This method may not seem to be a very short-cut, but, as applied in frequency tables and correlation tables, it will be found to be truly a short-cut.

8. Arithmetic Mean from Frequency Table.—Finding the arithmetic mean from a frequency table may be illustrated by use of Table XIII of freshman weights. The work is tabulated in Table XXXI. The usual assumption to make in regard to the distribution of the weights in any class is that the weights vary uniformly from item to item in the class. Under this assumption the average for any class would be the midpoint of the class. The sum of all the items in the class would be this midpoint, or average, multiplied by the frequency for that class, fX. The

size of item X in Table XXXI, is tabulated as the midpoints of the classes.

TABLE XXXI.—ARITHMETIC MEAN OF WEIGHTS OF FRESHMAN MEN IN POUNDS

Size of item X	Frequency f	fX
105	15	1,575
116	43	4,988
127	138	17,526
138	162	22,356
149	129	19,221
160	82	13,120
171	35	5,985
182	16	2,912
193	5	965
204	3	612
215	1	215
Total	$N = 629$	89,475

$$a = \frac{\Sigma X}{N} = \frac{89{,}475}{629} = 142.25.$$

9. Short-cut.—The work for the short-cut method is shown in Table XXXII. Let X' be chosen near the middle of the table so that the deviations $d_{X'}$, from X' will be as small as possible. Choose X' as one of the tabulated values of X so that $d_{X'}$, will be 0 for that class. Also, if X' is one of the tabulated values of X the deviations $d_{X'}$ will be multiples of the class interval, + for values of X greater than X', and − for values of X less than X'. It is also desirable that X' differ but little from a, so that $\Sigma d_{X'}$ shall be small.

TABLE XXXII.—ARITHMETIC MEAN OF WEIGHTS OF FRESHMAN MEN
SHORT-CUT METHOD

Size of item, in pounds X	Frequency f	$d_{X'}$	$fd_{X'}$
105	15	−55	− 825
116	43	−44	− 1,892
127	138	−33	− 4,554
138	162	−22	− 3,564
149	129	−11	− 1,419
160	82	0	0
171	35	11	385
182	16	22	352
193	5	33	165
204	3	44	132
215	1	55	55
Total.	N = 629		−11,165

$X' = 160.$

$$a = X' + \frac{\Sigma d_{X'}}{N}.$$

$$a = 160 + \frac{-11,165}{629},$$

$$a = 160 - 17.75,$$

$$a = 142.25.$$

10. Warning.—Do not make the mistake of adding the $d_{X'}$ column. This sum would give the sum of deviations for one item only from each class, or 11 items in all, instead of 629. There are 15 items with an average deviation of −55, making −55 × 15 = −825 for the sum of the deviations of these 15 items. Similarly, the sum of the deviations for the 43 items of the second class is −44 × 43 = −1,892, and so on down through the table. The sum of the deviations for the whole 629, that is, for the entire table and not for the $d_{X'}$ column alone, is

$$\Sigma d_{X'} = \Sigma(X - X') = -12,254 + 1,089 = -11,165.$$

This last sum may be written

$$\Sigma f d_{X'} = -11,165$$

if the context would thus be made clearer.

11. Another Example.—Table XXXIII is for finding the arithmetic mean of heights of freshmen. The data are obtained from Table XIV.

TABLE XXXIII.—ARITHMETIC MEAN OF HEIGHTS OF FRESHMAN MEN

Size of item, in pounds X	Frequency f	$d_{x'}$	$fd_{x'}$
60.5	3	−7.5	− 22.5
62.0	8	−6.0	− 48.0
63.5	33	−4.5	−148.5
65.0	51	−3.0	−153.0
66.5	115	−1.5	−172.5
68.0	156	0	0
69.5	148	1.5	222.0
71.0	64	3.0	192.0
72.5	43	4.5	193.5
74.0	5	6.0	30.0
75.5	3	7.5	22.5
Total.	$N = 629$		115.5

$X' = 68.0.$

$$a = X' + \frac{\Sigma d_{x'}}{N}$$

$$a = 68.0 + \frac{115.5}{629} = 68.0 + 0.18,$$

$$a = 68.18.$$

12. If a is Not One of the Given Items.—The arithmetic mean is frequently not one of the sizes of items in the universe represented. It may even be an impossible number. For example, the average size of family may be given as 5.63 persons. Of course, there could not be a fractional number of persons in a family. This need not destroy its value as an average.

13. Not Always Representative.—The arithmetic mean may not always be truly representative of the entire group of items. For example, suppose the following gives the height of trees in a dooryard.

HEIGHT OF TREES IN A DOORYARD

Height, in feet X	Number f	fX
40	2	80
50	5	250
60	2	120
300	1	300
	$N = 10$	750

$$a = \frac{750}{10} = 75 \text{ ft.}$$

If the statement is made that the 10 trees in this dooryard average 75 ft. high, someone seeing them might easily think that the average height was overstated. Seventy-five feet is not a truly representative height, as it could not be used as a single number describing the entire group of trees. Some other kind of average would doubtless be better.

14. Mode.—Another average is suggested by the remark: "Mr. A is a man of average height."

It is quite likely that the speaker means that Mr. A is of that height which occurs most frequently, the height, one may say, which is the style, *la mode.* Such an average is called the *mode.* The mode, then, may be defined as that *size which occurs most frequently.* If the items tend to concentrate about the mode, they are said to be stable or true to type.

The mode satisfies the general definition of an average, for if

$$X_1 = X_2 = X_3 = \ . \ . \ . \ = X_N = k,$$

the value which occurs most frequently is k, equal to each of the set.

A stricter definition may be given from consideration of a frequency distribution. *A mode is a value whose frequency is greater than the frequency of the smaller or of the larger sizes in the immediate neighborhood.* By this definition there might be more than one mode in a given set of values of a variable. An example may be seen in Table XII of freshman weights. The corresponding frequency polygon (Fig. 27) shows two quite distinct high points at weights 130.5 and 142.5 lb. These two modal weights may be modes for this sample of 629 freshmen only. It may be there is really only one modal weight for all freshmen. Some variables, however, do have more than one proper mode.

15. Mode Shown by Frequency Graph.—If there is a decided mode, a frequency table or frequency graph nearly always shows near what size of item it lies. Table XIII of weights of freshmen indicates a mode in class 133–143 lb. The smooth curve of Fig. 29 indicates the modal weight to be about 136 lb., this weight occurring at the ordinate of greatest length. Figure 31 seems to indicate a modal height of freshmen at about 68.25 in.

16.* Formula for Mode.—In order to derive a formula for computing the mode from a frequency table in which the modal class is well defined, examine Fig. 30, the rectangular histogram of freshman heights. The modal class (the class containing the mode) seems to be class 67.5–68.5. The height of the rectangle

of the next class to the right is greater than the height of the next rectangle to the left. It may be said, then, that the frequency o class 69.0 70.0 has greater weight in determining the mode tha does the frequency of class 66.0–67.0. The use of the wor "weight" suggests taking these two frequencies as weights an balancing them.

Let

f_1 = frequency of class next lower than the modal class.
f_2 = frequency of class next higher than the modal class.
l = lower boundary of the modal class.
c = class interval.
Z = mode.
x = distance from lower boundary of the modal class to wher the weights f_1 and f_2 balance.

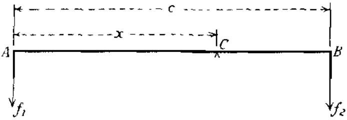


FIG. 34.

Hang weight f_1 at the lower boundary of the modal class an weight f_2 at the upper boundary of that class. The situation i seen in Fig. 34.

From 0 of the scale to A of the figure equals l. From 0 to l is the upper boundary of the modal class. $AB = c$, $AC = x$.

In order that the weights f_1 and f_2 shall balance at C

$$f_1 \times AC = f_2 \times CB,$$

or

$$f_1 x = f_2(c - x).$$

Solving for x gives

$$x = \frac{f_2 c}{f_1 + f_2}.$$

Then

$$Z = l + \frac{f_2 c}{f_1 + f_2}.$$

17. Illustrations.—Applying the formula in Table XIV:

$$f_1 = 115.$$
$$f_2 = 148.$$
$$l = 67.25.$$

$$c = 1.5.$$
$$x = \frac{148 \times 1.5}{115 + 148}.$$
$$x = \frac{222}{263}$$
$$x = 0.84.$$
$$Z = 67.25 + 0.84.$$
$$Z = 68.09.$$

This value of Z is less than that obtained graphically in Fig. 31. It is about 0.1 in. (1/15 of a class interval) greater than the midpoint of the modal class. This indicates a greater tendency to heights larger than the mid point of the modal class than to heights less than the midpoint of the modal class.

By the same formula compute the modal value of freshman weights from Table XIII. The class of greatest frequency is 133-143.

$$f_1 = 138.$$
$$f_2 = 129.$$
$$l = 132.5.$$
$$c = 11.$$
$$Z = 132.5 + x$$
$$= 132.5 + \frac{129 \times 11}{138 + 129}$$
$$= 132.5 + 5.3$$
$$= 137.8.$$

The mode, 137.8, is a little less than 138, the midpoint of the modal class. This could be foreseen, since in Fig. 28 the next rectangle to the left of the modal class is higher than the next to the right.

The mode is 0.2 lb. less than 138 lb., the midpoint of the modal class; 0.2 lb. is about two-hundredths of a class interval. This is much smaller than one-fifteenth, the deviation of the modal height from the midpoint of the modal class. This was to be expected since f_1 and f_2 are more nearly equal in the case of weights than in the case of heights.

18. Modal Class Not Well Defined.—If Table XII or Fig. 27 is used to determine modal weight of freshman men, the question at once arises as to whether the two modes indicated are apparent or real. From what was known of weights of men when this table and graph were made, there is no reason for thinking there should be two modes. This table serves very well for illustrating a

method of determining mode when the modal class is not well defined. This is known as the *method by grouping*.

TABLE XXXIV.—MODE BY GROUPING WEIGHT OF FRESHMEN
From Table XII

No. of class	Midpoint of class X	Frequency f	Groups I	II	III	IV	V	VI	VII	VIII	IX
1	110.5	15	27		49			83			
2	114.5	12		34		68			112		
3	118.5	22	56				100			162	
4	122.5	34		78	140						197
5	126.5	44	106			163		219			
6	130.5	62		119			175		236		
7	134.5	57	113		174					229	
8	138.5	56		117		172					214
9	142.5	61	116				158	198			
10	146.5	55		97	137				162		
11	150.5	42	82			107				137	
12	154.5	40		65			95				109
13	158.5	25	55		69			84			
14	162.5	30		44		59			72		
15	166.5	14	29				42			52	
16	170.5	15		28	38						
17	174.5	13	23								
18	178.5	10									

TABLE XXXV.—AUXILIARY TABLE

Group	Number of class							
	5	6	7	8	9	10	11	12
I					✓	✓		
II		✓	✓					
III			✓	✓	✓			
IV				✓	✓	✓		
V		✓	✓	✓				
VI	✓	✓	✓	✓				
VII		✓	✓	✓	✓			
VIII			✓	✓	✓	✓		
IX				✓	✓	✓	✓	
	1	4	6	7	6	4	1	

Mode at midpoint of class 8, $Z = 138.5$.

19. Mode by Grouping.—It has already been seen that increasing the class interval tends to smooth out irregularities due to the sampling. From Table XII the simplest way to increase the class interval is to double it by making a new classification whose first class includes the first two classes of Table XII, whose second class includes the third and fourth classes of Table XII, and so on. A new classification with double the class interval of Table XII may also be formed by making its first class include the second and third classes of Table XII, its second class to include the fourth and fifth classes of Table XII and so on. If both these new classifications are formed, it may happen that there is included only one modal class in each. If these modal classes overlap, it is reasonable to suppose that the class of Table XII which is common to both is the true modal class. If they do not overlap, either there is no distinct modal value, or else a further widening of the class interval is necessary to reveal it. The next step would then be to make new classifications with class interval three times that of the original table, Table XII. This may be done in three ways. The first class of the new classification may include the first three classes of the original table, it may include the second, third and fourth classes,

or it may include the third, fourth, and fifth classes of the original table. All three classifications should be formed. If each one develops a modal class and by inspection it is found that one of the classes of the original table is common to all three, it is quite likely that this is the modal class sought. It may happen, however, that in considering all five of the new classifications it will be found that there is not one and one only of the original classes occurring in these new classifications more times than does any other. This may make the modal class still uncertain. If so, four new classifications may be made in the same manner, each with a class interval equal to four times the class interval of the original table. This done, there are now formed nine classifications or frequency tables with class intervals greater than that of the original table. These nine groupings are examined as to their modal classes for the purpose of finding out how many times different classes of the original table occur in these modal classes. That class which occurs most often in these modal classes is considered to be the true modal class. Its midpoint is taken as the mode.

Usually it is not necessary to carry the process so far if there is really a single mode. Usually it will not be of any use to carry the process any farther if a modal class has not been found

20. The Method Applied.—The process may be condensed and tabulated as shown in Table XXXIV. The column headings are self-explanatory. It is evident that the first three and the last nine classes of Table XII will have no appreciable effect on the mode and so may be omitted from Table XXXIV.

21. Auxiliary Table.—Table XXXV is an auxiliary table for making it easy to pick out the class of the original classification which occurs most times in the modal classes of the new classifications. In following down the column of group I, the first new classification, a maximum frequency of 116 is found, which includes classes numbered 9 and 10 of the original classification. These two classes are then checked off in Table XXXV with group I. Similarly for group II, the maximum frequency of 119 includes classes 6 and 7 of the original classification. These are checked off with group II in Table XXXV. The checking of classes in the maximum frequencies for each of the groups is continued to the end of the table.

When the first two groupings are completed, it is found that the two classes of maximum frequency do not overlap. The first

includes classes 9 and 10 while the second includes classes 6 and 7. So no modal class is determined. When the next three groupings are completed and checked off in Table XXXV, it is at once seen that the class 8 is common to all three. It is quite likely then that class 8 is the modal class. If all five of the groupings are considered, however, it is seen that classes 7, 8, and 9 each occur in three of the maximum frequencies of these groupings. Classes 6 and 10 each occur in two of the maximum frequencies. It still seems likely that the mode is the midpoint of class 8. By continuing the process of new classifications to form more groupings and by checking the classes of maximum frequency in Table XXXV, it is found that class 8 occurs in 7 of the maxima. This is a greater number of times than any other class occurs. It is then concluded that class 8 is the modal class and its midpoint, 138.5, is regarded as the mode. Note the totals of number of times classes 5 to 11 inclusive occurred in the maximum frequencies of the groups. They are 1, 4, 6, 7, 6, 4, 1, shown in the last line of Table XXXV. The symmetry of arrangement of these numbers strengthens the conclusion that the midpoint of class 8 is the mode. Such symmetry could not be expected to always occur.

There is no doubt now that the two modes displayed in Fig. 27 were apparent only. In drawing the smooth curve to show general tendencies the two high points should be smoothed out.

22. Mode from the Ogive.—*The mode may be determined* graphically from the ogive by determining where the curve is *steepest. At that point drop a perpendicular to the base line.* The scale mark where this perpendicular strikes the base line is the value of the mode. The steep portion of the curve is usually so nearly straight that a single point where steepness is greatest is not easy to determine. For this reason this is not a very satisfactory way of determining the mode. It may be done fairly well by beginning on the lower part of the curve, moving a straight edge along the curve, and keeping it tangent to the curve. It rotates counterclockwise. Mark the point on the curve where the tangent ceases to rotate. Then do the same way, beginning on the upper part of the curve, locating another point where the tangent ceases to rotate. Halfway between the two points should be close to the point where the curve is steepest. Doing this in Fig. 32 gives the mode as about 140 lb.

23. Comparison of Results.—Compare, now, the values of the mode as obtained by the different methods.

Graph (Fig. 28), about 137.5.
Formula (Table XIII), 137.8.
Grouping (Table XXXIV), 138.5.
Ogive (Fig. 32), 140.

24. Effect of Shifting Class Limits.—The first three of these values are not very different. The question may arise as to which is the best value. That at once brings up the question as to what difference it would make if the class limits of Table XIII were shifted. With a class interval of 11, there are 11 different possible numbers with which to begin the first class. One, for example, would be to make the limits of the first class 93 to 103.

Turning to Table II, Chap. I, the resulting frequency table can at once be made from the first and last columns. The result is Table XXXVI.

TABLE XXXVI.—WEIGHTS OF FRESHMAN MEN

Classes in pounds X	Frequency f
93–103	3
104–114	21
115–125	84
126–136	145
137–147	159
148–158	108
159–169	60
170–180	34
181–191	8
192–202	4
203–213	2
214–224	1
	$N = 629$

The modal class is indicated as 137–147.

$$Z = 136.5 + \frac{11 \times 108}{145 + 108} = 136.5 + 4.7 = 141.2.$$

This result is considerably larger than any of the values previously found. The mode, then, may be affected by the position

of class limits. What position of class limits shall be used? By tabulating and plotting frequency polygons for five different positions of class limits, ranging over the different possible positions, it was judged that the classes as given in Table XIII seemed to be most fairly representative. The mode was computed for each one. The values ranged from 135.9 to 142.0. The average of the five values was found to be 139.3.

25. Effect of Position of Class Limits in Grouping.—Treating each group of Table XXXIV as a frequency table and computing the mode by formula, the values range from 133.3 to 143.9 with an average of 138.7. This average is very close to the value obtained by the method of grouping, as would be expected. The value obtained by grouping, being an approximate average of results from a systematic group of frequency distributions, would seem to be a reasonable result.

The extreme values came from groups I and II. Examination of Table XXXV would show this. Neither of these groupings would have given a fair representative frequency table for use in statistical analysis.

Having once decided on the best position for class limits, if a decided modal class appears, the method by formula or by graph would ordinarily be used.

26. Effect on Arithmetic Mean.—At this point it may be of interest to see whether the position of class limits of Table XXXVI makes much difference in the arithmetic mean as computed from the frequency table. The computation can be carried out by adding an fX column. If this is done, ΣX is found to be 89,461, giving the arithmetic mean as 142.24. From Table XXXI it was found to be 142.25. There is no appreciable difference. This justifies the assumption in this case that the midpoint of a class may be used as the average of the items in that class.

27. Median.—If a military company is lined up, company front, with the shortest man at one end, the next taller next, and so on up to the tallest man at the other end, the men are *arrayed* as to height. The height of the man at the middle of the array is the *median height* of the men in the company. It is called the *median* man height of the men in the company. If there is an even number of men in the company, there is no middle man. Then the median may be considered as the arithmetic mean of the heights of the two middle men. It is a comparatively easy matter to array, count up, and find the median of a set of

values of a variable, provided the total number N of values is not large.

28. More Precise Definition.—If N is large, the items may be tabulated in a frequency table, and that class between whose boundaries the median lies is easily determined from the column of cumulative frequencies. For example, use Table XIII of weights of freshman men.

$$N = 629.$$

$$\frac{N}{2} = 314.5.$$

From the column of cumulative frequencies it is seen that the first three classes contain 196 items, and the first four classes contain 358 items.

$$196 < 314.5 < 358.$$

The median lies, therefore, between the boundaries of the fourth class,

$$132.5 < M < 143.5.$$

Think of all the items of the table as being put in array, each item occupying a unit segment of the base line on which it stands. At points on the base line whose abscissas are equal to the cumulative frequencies, plot the respective values of the class boundaries as ordinates. Connect the top of these ordinates by straight lines from each to the next, forming a graph of the array. Then on the assumption that within any class there is uniform variation of size of item throughout that class, the ordinate to the graph at the midpoint of any unit segment of the base line will measure the size of item standing on that unit segment. The entire length of base line occupied by the array equals N. The point halfway through the array will be at the scale mark $N/2$. The value of the ordinate at this point is *defined* as the *median*.

For a small number of items, such as the man heights of the military company, ordinates may be drawn at the midpoints of unit segments of the base line representing each man height, and a smooth curve drawn through the tops of these ordinates. The *median is the value of the ordinate to this curve at the midpoint* of the base line used. As an illustrative example, in Fig. 35 are arrayed the heights of a squad of 16 men selected at random throughout the company. The 16 ordinates measuring the heights occupy 16 units of base line. A curve is drawn through the tops of these ordinates. Halfway through the array at scale

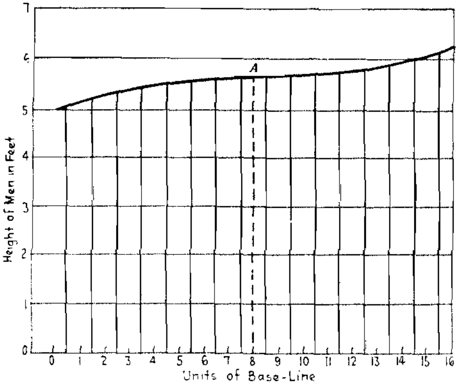


Fig. 35.—Array of heights of sixteen soldiers.

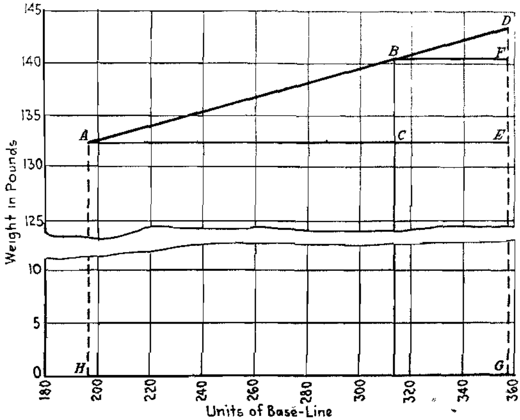


Fig. 36.—Graph of median class of freshman weights for determining formula for the median.

mark 8, the ordinate $8A$ is drawn. Its length, 5.7 ft., is the value of the median man height of the squad.

29. Application to Frequency Table.—For the median freshman weight, it is not necessary to graph the entire array. Only the class in which the median lies need be used. As has just been seen, the lower boundary of this class is 132.5 and the upper boundary, 143.5. The number in the class is 162. Proper scales are selected and the graph drawn for this class only (Fig. 36). There are 196 items up to this class. An ordinate equal to 132.5, the lower boundary of the class, is drawn at 196 on the base line. An ordinate equal to 143.5, the upper boundary of the class, is drawn at 358 on the base line. The upper ends of these ordinates are connected by the straight line AD. At 314.5, the point halfway through the entire array, the ordinate measuring the median is drawn, meeting the line AD at B. Through A a horizontal line is drawn cutting the ordinates to B and D at the points C and E, respectively. Through B a horizontal line is drawn meeting the ordinate to D at F.

30. Interpolation Formula for Median.—Use the following notation:

M = median.

N = number of items in entire array, entire length of base line used.

$l = HA$ = lower boundary of the class used.

$l' = GD$ = upper boundary of the class used.

$c = ED$ = class interval.

$f = AE$ = *frequency of the class.*

F_M = cumulative frequency up to the class.

F'_M = cumulative frequency down to the class from the upper end of the array.

$i = AC = N/2 - F_M$ = number of units of base line that must be passed from the lower boundary of the class to reach the median.

$i' = BF = N/2 - F'_M$ = number of units of base line that must be passed from the upper boundary of the class to reach the median.

$x = CB$ = amount to be added to l to give M.

$x' = FD$ = *amount to be subtracted from l' to give M.*

From the frequency table:

$N = 629.$

$$l' = 143.5.$$
$$c = 11.$$
$$f = 162.$$
$$F_M = 196.$$
$$F'_M = 271.$$
$$i = \frac{629}{2} - 196 = 314.5 - 196 = 118.5.$$
$$i' = \frac{629}{2} - 271 = 314.5 - 271 = 43.5.$$

From the similar triangles ACB and AED in the figure,

$$\frac{CB}{ED} = \frac{AC}{AE}.$$

Substituting from the above notation

$$\frac{x}{c} = \frac{i}{f},$$
$$x = \frac{ci}{f}.$$
$$M = l + x,$$
$$M = l + \frac{ci}{f}.$$

This is the formula for the median interpolating from the lower boundary up.

Also from the similar triangles BFD and AED

$$\frac{FD}{ED} = \frac{BF}{AE};$$

whence,

$$\frac{x'}{c} = \frac{i'}{f},$$
$$x' = \frac{ci'}{f},$$

and

$$M = l' - x',$$

or

$$M = l' - \frac{ci'}{f},$$

the formula for the median interpolating from the upper boundary down.

Using the first formula and the values from the frequency table,

$$M = 132.5 + \frac{11 \times 118.5}{162},$$
$$M = 132.5 + 8.05,$$
$$M = 140.55.$$

Or, from the second formula,

$$M = 143.5 - \frac{11 \times 43.5}{162},$$
$$M = 143.5 - 2.95,$$
$$M = 140.55.$$

The two formulas, which are consistent, must give the same result. Use of both gives a check on the work.

31. Example.—The median freshman height is computed from Table XIV.

$N = 629.$
$l = 67.25.$
$l' = 68.75.$
$c = 1.5.$
$f = 156.$
$F_M = 210.$
$F_M' = 263.$
$i = 314.5 - 210 = 104.5.$
$i' = 314.5 - 263 = 51.5.$

Then

$$M = 67.25 + \frac{1.5 \times 104.5}{156},$$
$$M = 67.25 + 1.01,$$
$$M = 68.26,$$

or

$$M = 68.75 - \frac{1.5 \times 51.5}{156},$$
$$M = 68.75 - 0.49,$$
$$M = 68.26.$$

32. Median from the Ogive.—The median may be at once determined graphically from the ogive by bisecting the total height of the curve. From the point of bisection draw a horizontal line to the curve. From the point where it meets the curve, drop a vertical line to the base line. The reading of the horizontal scale at this point is the value of the median. In this manner

from Fig. 32 the median freshman weight is found to be about 140¾ lb. From Fig. 33 the median height is about 68¼ in. These results are very close to the values that were computed by interpolation formula. These were 140.50 lb. and 68.26 in.

33. Theorem.—If a set of values of a variable be placed in array and the median is taken as that value which stands in the middle of the array, the following theorem may be proved:

The sum of differences from the median, dropping algebraic signs, is a minimum.

This means that if the differences between the median and each of the other items be added, the sum will be less than the sum of the differences from any other item.

Suppose the items $OM_1, OM_2, \ldots OM_N$,

$O \quad M_1\ M_2\ M_3 \quad M_4\ M_5 \quad M_6 \quad M_7 \quad M_N$

all laid off from a common origin O. Choose any one, OM_4, from which to take differences. Take any two items, one less than OM_4 and one greater than OM_4, say OM_2 and OM_7.

$$OM_4 - OM_2 = M_2M_4$$
$$OM_7 - OM_4 = M_4M_7$$

adding, $$M_2M_4 + M_4M_7 = M_2M_7 = OM_7 - OM_2.$$

The sum of these two differences from OM_4 equals the difference of the two items taken.

Take any two items each less than OM_4, say OM_1 and OM_3.

$$OM_4 - OM_1 = M_1M_4,$$
$$OM_4 - OM_3 = M_3M_4.$$

Since

$$M_1M_4 > M_1M_3,$$
$$M_1M_4 + M_3M_4 > M_1M_3.$$

But

$$M_1M_3 = OM_3 - OM_1.$$

The sum of these two differences, therefore, is greater than the difference of the two items taken.

Take any two items each greater than OM_4, say OM_5 and OM_6.

$$OM_5 - OM_4 = M_4M_5,$$
$$OM_6 - OM_4 = M_4M_6.$$

Since

$$M_4M_6 > M_5M_6,$$
$$M_4M_5 + M_4M_6 > M_5M_6.$$

But

$$M_5M_6 = OM_6 - OM_5.$$

The sum of these two differences, therefore, is greater than the difference of the two items taken. It is seen, then, that if the items are taken in pairs, one item in each pair being less than OM_4 and the other greater than OM_4, the sum of all differences will be less than if in some pairs both items are less than OM_4 or both greater than OM_4. So in order to make the sum of the differences a minimum, OM_4 must be chosen so that all the other items may be paired off with one item of each pair less than OM_4 and the other greater than OM_4. The number of items less than OM_4 must, then, equal the number of items greater than OM_4. In other words, OM_4 must stand in the middle of the array and is the median.

If there are an even number of items in the array, any value between the values of the two middle items will satisfy the theorem. The most reasonable value to use in this case is the arithmetic mean of these two items.

The same theorem applies to the more precise definition of the median as an ordinate. If N is odd, the ordinate measuring the median will coincide with the middle item of the array, and if N is even, it will fall halfway between the two middle items.

34. Median is an Average.—The median satisfies the general definition of an average.

$$\text{If } X_1 = X_2 = X_3 = \ . \ . \ . \ = X_N = k,$$

and the items be put in array with a line drawn connecting their tops, the line will be straight, and parallel to the base line. The ordinate at the middle of the array will equal k. Therefore, the median equals k, and equals each one of the items.

35. Three Averages.—The three constants, arithmetic mean, mode, and median, are of frequent use as averages in statistical analysis. They are each used as type forms, or central values around which all the values of the variable group themselves more or less closely.

No matter how carefully samples may be selected from a given universe, the average is quite likely to vary somewhat from sample

to sample. That form of average would be considered most representative for which the differences from sample to sample are least.

36. Each a Type Form.—The sum of the deviations from the arithmetic mean is zero, warranting its use as a central type form. *The sum of the differences between the median and each of the items is a minimum.* This makes the median what may be called the size of item of closest fit to the entire set. It is thus a central type form. The mode is the size of item of most frequent occurrence, which makes it also a type form.

37. Median from Frequency Curve.—Since the area under the frequency curve from the left end up to any ordinate is proportional to the total frequency up to that ordinate, it follows that the ordinate which bisects the area stands on the base line at the value of the median. *In other words, the ordinate at the median passes through the center of area.*

38. Arithmetic Mean from Frequency Curve.—Take Fig. 28, the rectangular histogram of freshman weights, and note the geometric interpretation of the process of computing arithmetic mean as shown in Table XXXI. If the width of each rectangle is regarded as a unit, then the values of f in the table represent the areas of the corresponding rectangles. The values of X are the distances of the centers of the rectangles from the vertical axis of reference. *$N = \Sigma f$, the total frequency, is the entire area.* fX for any rectangle equals the moment of its area about the vertical axis.

$$a = \frac{\Sigma fX}{\Sigma f}.$$

From physics it is known that the sum of the moments of the rectangles about any axis, divided by the sum of all the rectangles, gives the distance of the center of gravity of the entire area from that axis. **It follows that the ordinate at the point *a* of the base line goes through the center of gravity of the area. Since the area under the smooth curve is the limit of area of the sum of the rectangles as the width of the rectangles becomes indefinitely small, the ordinate through the center of gravity of the smooth curve should be at the arithmetic mean of the universe of values from which the sample was taken.**

The arithmetic mean and the median are then central values from a geometric standpoint.

Geometrically, the mode is located at the point where the greatest ordinate to the curve stands, the ordinate where the tangent to the curve is parallel to the base line.

39. Geometric Properties Serve as Checks.—These geometric properties of the arithmetic mean and the median and mode may be used to check the correctness of smoothing the frequency diagram. See if the ordinate at M bisects the area under the curve, and if the ordinate at a goes through the center of gravity of the area. This latter may be ascertained by drawing the curve on stiff paper or cardboard. Cut out the figure drawn. Lay it on a knife edge with the base line perpendicular to the knife edge, and find where it balances. At this point on the base line read the value of the arithmetic average as given by the curve. See if it checks with the computed value. See if the greatest ordinate to the curve stands on the base line at the computed value of the mode.

40. One Average May Be More Representative than Another. *Sometimes one average is more representative than another.* It was shown on page 102 that the arithmetic mean of the height of trees in a dooryard was not truly representative. From the table of heights it is at once seen that the median is 50 as is also the mode. This value of 50 is more representative of these tree heights than is 75, the arithmetic mean.

41. Advantages and Disadvantages of Each.—The size of the arithmetic mean is influenced by the sizes of extreme items. The median is not influenced by extreme items except in so far *as items of extreme size be added to the set of items. Since* N is then changed, the position of the median in the array is changed. The items in the central portion of the array are usually all of nearly the same size. So the size of the median would be changed but little if any. The mode is not at all affected by extreme items unless many of them are introduced.

The arithmetic mean and the median can be determined quite closely by formula. Sometimes it is difficult to determine a mode. The arithmetic mean and the median always exist. Sometimes there is no mode at all.

The product of arithmetic mean and number of items gives the totality of all items. There is no such simple relation for either mode or median.

The median is not so much affected by accidental variations as is the arithmetic mean. A low median wage almost surely means

a group of low wages. Suppose, for example, in a certain shop there are 100 men at wages of $2 per day, and three men getting $100 each per day. The arithmetic mean wage is nearly $5 per day, not a small wage. The median wage is $2, which is more representative of wages in that shop, and indicates a low wage scale.

There are some characteristics of certain variables that are hard to measure numerically and yet the variates may be ranked, or arrayed, with respect to the given characteristic. In such a case the arithmetic mean cannot be found, but the variate having the median characteristic is determinable.

It sometimes happens that either the median or the arithmetic mean may occur at a point in the distribution where there are few or no items. If there is a mode, it may then be a better representative type form.

If the frequency curve has a symmetric bell shape, the arithmetic mean, the mode, and the median are equal to each other. This can be seen at once from the frequency curve and the geometric properties above stated.

42. Weighted Arithmetic Mean.—A weighted arithmetic mean is an arithmetic mean obtained when each of the items or some of them have been counted more than once.

In Table XXXI the column of "size of item" contains 11 numbers. In obtaining the arithmetic mean, the first size, 105, was counted 15 times, as shown by the corresponding frequency. The second size of 116 was counted 43 times, and so on down the table. In other words, the size at the midpoint of each class was weighted by the corresponding class frequency. To get the arithmetic mean the products fX, of size by weight, were added, and the sum divided by N, the sum of the weights. The result is the weighted arithmetic mean of the sizes X in the first column of the table.

43. Rule.—The process gives the rule for finding weighted arithmetic mean. Multiply each value of the variable by its weight, add the products, and divide by the sum of the weights.

If $X_1, X_2, X_3, \ldots X_N$ are the values of the variable and $w_1, w_2, w_3, \ldots w_N$ are the corresponding weights, then the weighted arithmetic mean equals

$$\frac{w_1X_1 + w_2X_2 + w_3X_3 + \ldots + w_NX_N}{w_1 + w_2 + w_3 + \ldots + w_N} = \frac{\Sigma wX}{}$$

44. Weighted Observations.—Sometimes in making a set of observations, certain ones of them are more reliable or accurate than others. Then in getting the average of the set, weights are given to the different observations according to their reliability and a weighted mean found. For example, a surveyor might measure the length of a line by pacing, by using a yardstick, or by using a steel tape. The yardstick measurements, if done carefully, should have greater weight than the results by pacing. The steel-tape measurements should be much better than those by either of the other methods and so be given much greater weight. When the weights are decided upon, a weighted arithmetic mean of all measurements would be used as the most probable value of the length of the line.

45. Weighted Grades.—When a student is graded in a certain subject, his final examination grade may be given a weight of 1, his recitation work a weight of 2, and his laboratory work a weight of 3. Suppose his final examination mark is 60, his recitation grade is 80, and his laboratory work is 95. His recorded grade will be the weighted arithmetic mean,

$$\frac{60 \times 1 + 80 \times 2 + 95 \times 3}{1 + 2 + 3}$$
$$= \frac{505}{6}$$
$$= 84\tfrac{1}{6}.$$

The grade of 95 is not as much above 80 as 60 is below it; yet its greater weight brings the final average considerably above 80.

46. Geometric Mean.—Another average of not uncommon use is the *geometric mean.* The geometric mean of n quantities is the nth root of their product. The geometric mean of X_1, X_2, X_3, . . . X_N is

$$\sqrt[N]{X_1X_2X_3 \ . \ . \ . \ X_N}.$$

The G. M. of 4 and 9 is

$$\sqrt{4 \times 9} = 6.$$

The G. M. of 2, 4, and 8 is

$$\sqrt[3]{2 \times 4 \times 8} = 4.$$

The G. M. of 2, 4, 8, 16, and 32 is

$$\sqrt[5]{2 \times 4 \times 8 \times 16 \times 32} = 8.$$

Write the series 2, 4, 8, 16, 32

as $2^1, 2^2, 2^3, 2^4, 2^5$.

This is a geometric progression with common ratio of 2. The exponents of 2 form an arithmetic progression with common difference of 1. In general, a geometric progression of n terms is written

$$a, ar, ar^2, ar^3, \ldots ar^{n-1}.$$

The exponents of r,

$$0, 1, 2, 3, \ldots n-1,$$

always form an arithmetic progression.

The G. M. of $X_1, X_2, X_3 \ldots X_N$ is

$$\sqrt[N]{X_1 X_2 X_3 \ldots X_N}.$$

$$\text{Log G. M.} = \frac{\log X_1 + \log X_2 + \log X_3 + \ldots + \log X_N}{N}.$$

That is, the logarithm of the G. M. of a set of numbers is the A. M. of their logarithms. This makes a ready method of computing the G. M. of a set of numbers. Take the arithmetic mean of their logarithms and look up its antilogarithm.

47. Connection between G. M. and A. M.—These considerations show an intimate connection between geometric mean and arithmetic mean. To any property of the G. M. there is a corresponding property of the A. M. The G. M. applies to a series of terms that are to be multiplied together, as A. M. applies to a series of terms to be added together. G. M. may be applied to interpolation when there is a uniform rate of change, and A. M. when there is a uniform absolute change. G. M. applies to compound interest, as A. M. applies to simple interest; G. M. to a series of relative prices, and A. M. to a series of price increases.

48. Illustration from Interest.—The compound amount of \$1 at 5 per cent compounded annually for 10 years is $(1.05)^{10} = \$1.63$. The compound amount for 20 years is $(1.05)^{20} = \$2.65$. The compound amount for 15 years, the middle of the period from 10 to 20 years, is G. M. of \$1.63 and \$2.65

$$\begin{aligned} &= \sqrt{1.63 \times 2.65} \\ &= \sqrt{4.3195} \\ &= 2.08 \\ &= (1.05)^{15}. \end{aligned}$$

G. M. of $(1.05)^{10}$ and $(1.05)^{20}$ is

$$\begin{aligned} &\sqrt{(1.05)^{10} \times (1.05)^{20}} \\ &= \sqrt{(1.05)^{30}} \\ &= (1.05)^{15}. \end{aligned}$$

The simple amount of \$1 at 5 per cent per annum for 10 years is

$$1 + 10 \times .05 = 1.50.$$

For 20 years,

$$1 + 20 \times .05 = 2.00.$$

The A. M. of \$1.50 and \$2 gives the simple amount for 15 years, the middle of the period from 10 to 20 years.

$$\frac{1.50 + 2.00}{2} = 1.75.$$

Simple amount for 15 years is

$$1 + 15 \times .05 = 1.75.$$

49. Average Rate of Growth.—Suppose the population of a certain city is 100,000 in 1890. The census of 1900 shows a $12\frac{1}{2}$ per cent increase. The census of 1910 shows a 50 per cent increase. The census of 1920 shows a 100 per cent increase. What has been the average per cent increase per decade?

The average per cent increase applied each decade would give the 1920 population the same as given by the above census figures. The census figures give

Date	Population
1890	100,000
1900	112,500
1910	168,750
1920	337,500

$$100{,}000(1.125 \times 1.50 \times 2.00) = 337{,}500.$$

Let x per cent be the average rate, then

$$100{,}000\{(1 + x) \times (1 + x) \times (1 + x)\} = 337{,}500,$$
$$100{,}000(1 + x)^3 = 337{,}500.$$

Then

$$(1 + x)^3 = 1.125 \times 1.50 \times 2.00,$$
$$1 + x = \sqrt[3]{1.125 \times 1.50 \times 2.00},$$
$$1 + x = \sqrt[3]{3.375},$$
$$1 + x = 1.50,$$
$$x = .50 = 50 \text{ per cent}.$$

The required average per cent increase is $x = 50$ per cent per decade. It is the G. M. of 1.125, 1.50, and 2.00.

50. Interpolation.—Suppose the population of a certain city in 1910 was 100,000 and in 1920, 144,000. What [illegible]e population in 1915?

If the population increased the same number of people each year, the population in 1915 was the A. M. of 100,000 and 144,000, or

$$\frac{100{,}000 + 144{,}000}{2} = 122{,}000.$$

If the population each year increased at the same per cent of the population at the beginning of the year, then the population in 1915 was the G. M. of 100,000 and 144,000, or

$$\sqrt{100{,}000 \times 144{,}000} = 120{,}000.$$

The two ways of increase are shown graphically in Fig. 37. The straight line *ABC* is the population graph if the change

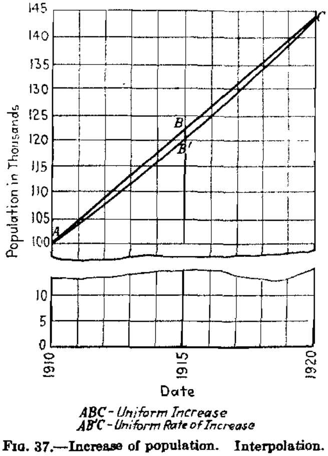


FIG. 37.—Increase of population. Interpolation.

from year to year was the same amount. The curve *AB′C* is the population graph if the per cent increase was the same each year. The ordinates to *ABC* at the different dates form an arithmetic, or constant difference, progression. The ordinates to *AB′C* at the different dates form a geometric, constant-ratio progression. In dealing with differences, an arithmetic mean is likely to be useful; in dealing with ratios, a geometric mean.

51. G. M. is an Average.—The geometric mean satisfies the general definition of an average.

If

$$X_1 = X_2 = X_3 = \ . \ . \ . \ = X_n = k,$$

$$\text{G.M.} = \sqrt[n]{X_1 X_2 X_3 \ . \ . \ . \ X_n} = \sqrt[n]{k^n} = k,$$

equals each of the items.

52. Harmonic Mean.—Another kind of average occasionally used is the *harmonic mean.* The harmonic mean of a series of numbers is the reciprocal of the arithmetic mean of their reciprocals. In algebraic symbols,

$$\text{H.M. of } X_1, X_2, X_3, \ . \ . \ . \ X_N$$

is

$$\frac{1}{\dfrac{\dfrac{1}{X_1} + \dfrac{1}{X_2} + \dfrac{1}{X_3} + \ . \ . \ . \ + \dfrac{1}{X_N}}{N}}.$$

This reduces to

$$\frac{N}{\sum \frac{1}{X}}.$$

The harmonic mean of a set of items may be defined as the number of items divided by the sum of their reciprocals.

53. Illustrations.—Suppose a man in traveling 3 miles goes the first mile at the rate of 8 miles per hour, the second mile at the rate of 10 miles per hour, and the third mile at the rate of 12 miles per hour. What is his average speed in miles per hour?

The first mile takes $\frac{1}{8}$ hr.,
The second mile takes $\frac{1}{10}$ hr.,
The third mile takes $\frac{1}{12}$ hr.,
The three miles take $\frac{1}{8} + \frac{1}{10} + \frac{1}{12} = \frac{37}{120}$ hr.
Then the average time per mile is

$$\tfrac{1}{3} \text{ of } \tfrac{37}{120} = \tfrac{37}{360} \text{ hr.}$$

So the average miles per hour is $\frac{360}{37} = 9.73$.

It is seen that the reciprocals of 8, 10, and 12 were added, the sum divided by 3, the number of items, and the reciprocal of the result taken. In other words, the average speed in miles per hour is the harmonic mean of 8, 10, and 12.

Suppose sugar is 12 lb. for \$1, flour, 20 lb. for \$1, and potatoes, 25 lb. for \$1, and it is necessary to find the average price per

pound. If the average price is obtained by using the arithmetic mean of 12, 20, and 25, the result will be the harmonic mean of prices. Thus:

$$\frac{12 + 20 + 25}{3} = 19$$

The A. M. of the number of pounds for \$1 is 19; 19 lb. for \$1 is \$1/19 per pound.

The prices per pound are: sugar, $8\frac{1}{3}$ cts., flour, 5 cts., potatoes, 4 cts.

The H. M. of these prices is

$$\frac{3}{\frac{3}{25} + \frac{1}{5} + \frac{1}{4}} = \frac{300}{57} = \frac{100}{19}\text{ cts.} = \$\frac{1}{19}.$$
$$= 5\tfrac{5}{19}\text{ cts.}$$

The A. M. of prices per pound is

$$\frac{\frac{1}{12} + \frac{1}{20} + \frac{1}{25}}{3} = \$\frac{52}{900}$$
$$= 5\tfrac{7}{9}\text{ cts.}$$

If this method is used to get the average number of pounds for \$1, the result will be:

$$900/52 = 17\tfrac{4}{13}\text{ lb. for \$1.}$$

This is the H. M. of 12, 20, and 25, the respective number of pounds for \$1.

54. Relations between A. M., G. M., and H. M.—The *arithmetic mean, geometric mean, and harmonic mean are related mathematically.* If only two numbers, X_1 and X_2, are involved, their G. M. is the geometric mean of their A. M. and H. M.

A. M. of X_1 and X_2 is $\dfrac{X_1 + X_2}{2}$.

H. M. of X_1 and X_2 is $\dfrac{2}{\frac{1}{X_1} + \frac{1}{X_2}} = \dfrac{2X_1X_2}{X_1 + X_2}$.

Geometric mean of A. M. and H. M. is

$$\sqrt{\frac{X_1 + X_2}{2} \cdot \frac{2X_1X_2}{X_1 + X_2}} = \sqrt{X_1X_2}$$
$$= \text{G. M. of } X_1 \text{ and } X_2.$$

It has been proved that in any series of positive numbers

H. M. < G. M. < A. M.

55. H. M. is an Average.—The harmonic mean satisfies the general definition of average. If

$$X_1 = X_2 = X_3 = \ . \ . \ . \ = X_N = k,$$

$$\text{H. M.} = \frac{N}{\frac{1}{X_1} + \frac{1}{X_2} + \frac{1}{X_3} + \ . \ . \ . \ + \frac{1}{X_N}} = \frac{N}{\frac{N}{k}} = k,$$

equal to each of the items.

Arithmetic mean, geometric mean, and harmonic mean are all used in formulas for index numbers.[1]

56. Other Averages.—Besides these averages an infinite number of different averages satisfying the general definition might be devised. The square root of the arithmetic mean of the squares of a set of numbers is an example,

$$\sqrt{\frac{\Sigma X^2}{N}}.$$

It is called the root-mean-square of the set of numbers. In the next chapter the root-mean-square of deviations from A. M., called standard deviation, will be introduced.

Exercises

1. Find the arithmetic mean of the following numbers and check that the sum of their deviations from the mean is zero:

$$17, 43, 21, 18, 104, 62, 47, 96, 117, 8.$$

2. Find the deviations of the numbers in Ex. 1 from 50 and check that their arithmetic mean added to 50 gives the arithmetic mean of the numbers themselves.

3. Using the frequency table of mean monthly temperatures that was made from Table IV in Ex. 2, Chap. II, find the arithmetic mean monthly temperature, by the direct method, for the period of time covered.

4. Find the same mean monthly temperature by the short-cut method.

5. Using the frequency table of monthly precipitation that was made from Table III in Ex. 3, Chap. II, find the arithmetic mean monthly precipitation, by the direct method for the period of time covered.

6. Find the same mean monthly precipitation by the short-cut method.

7. Find by formula the modal monthly temperature, using the table made in Ex. 2, Chap. II. *How much does it differ from the arithmetic mean?*

8. Find by formula the modal monthly precipitation, using the table made in Ex. 3, Chap. II. How much does it differ from the arithmetic mean?

9. Find the median monthly temperature, using the table made in Ex. 2, Chap. II.

[1] Prof. Irving Fisher, "The Making of Index Numbers."

10. Find the median monthly precipitation, using the table made in Ex. 3, Chap. II.

11. Check that the sum of the differences of the following set of numbers from their median is less than the sum of the differences from some other number chosen at random:

47, 63, 87, 99, 21, 27, 35, 38, 55.

12. The arithmetic mean of a set of 20 numbers is 75. What is the sum of these numbers?

13. In chemistry a student was graded 60 in final examination, 90 in recitation, and 80 in laboratory. These grades were weighted 1, 2, and 3, respectively. Find the student's average grade. If the weights had been 1, 3, and 6, respectively, what change would it have made in the final mark? What if the weights had been 1, 5, and 10?

14. From the table of population of the state of Washington, given in Ex. 10, Chap. III, determine the average rate of increase per decade for the period covered by the table.

15. Determine the average increase in population per decade in the state of Washington for the same period of time.

16. The number of bacteria in a certain culture was found to be 4×10^6 at noon of one day. At noon the next day the number was found to be 9×10^6. If the number increased at a constant rate per hour, how many bacteria were there at midnight?

17. If a man travels 6 miles at a speed of 4 miles per hour for the first 3 miles, 3 miles per hour for the next 2 miles, and 2 miles per hour for the last mile, what is his average speed in miles per hour?

18. A man has $10,000 invested at 5 per cent simple interest, $5,000 at 6 per cent, and $1,000 at 10 per cent. What is his average rate of interest?

19. What relation must exist between a and b in order that their arithmetic mean, geometric mean, and harmonic mean may all be equal?

20. At the end of 5 years a certain sum of money placed at 6 per cent per annum, compounded annually, amounts to $1,338.23. At the end of 15 years it amounts to $2,396.56. What is the amount at the end of 10 years?

21. At the end of his first quarter in college a freshman had credits as follows: 5 hours of trigonometry with a grade of 97, 4 hours of English with a grade of 88, 3 hours of history with a grade of 85, and 4 hours of physics with a grade of 89. What was his average grade per hour of credit?

22. In a certain park there are 4 alder trees each 45 ft. high, 10 maples each 50 ft. high, 2 hemlocks each 200 ft. high, and 4 Douglas firs each 250 ft. high. What is the average tree height of the trees in this park? What kind of average do you think is fairest to use? Is it fair to group these tree heights in one group?

23. The population of a certain city increased in 10 years from 335,000 to 495,000. What was the average increase per year? What was the average annual rate of increase?

24. Find from Table XXIII the arithmetic mean, mode, and median leaf length. Arrange them in order of size.

25. Find from Table XXIV the arithmetic mean, mode, and median leaf breadth. Arrange them in order of size.

26. In Table XV what was the average number of heads per throw? Is the arithmetic mean a good average to use? What is the significance of the fractional value in the arithmetic mean?

27. From the table made in Ex. 19, Chap. II, determine the arithmetic mean, mode, and median leaf length.

28. From the table made in Ex. 20, Chap. II, determine the arithmetic mean, mode, and median leaf breadth.

CHAPTER V

DISPERSION

Two frequency distributions may be alike in every respect except that the average of one is greater than the average of the other. Such are the distributions shown by the curves in Figs. 38 and 39. Such distributions may be compared by comparing their averages.

Two frequency distributions may have exactly the same average, but the variation of one with respect to the average may

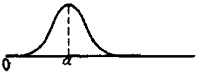


FIG. 38.

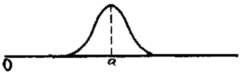


FIG. 39.

be quite different from that of the other. The two curves in Fig. 40 show such distributions. For many purposes sufficient information is not given for comparison of these variables by saying that they have the same average. Some measure of the variation should be devised by means of which the two variables may be compared.

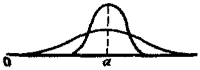


FIG. 40.

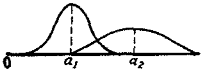


FIG. 41.

Two frequency distributions may have different averages and also different variation with respect to the average. The curves of Fig. 41 show two such distributions. In comparing these two variables both the average and the measure of variation must be considered.

1. Dispersion.—It thus appears that, in order to compare two variables, the type form or average is not sufficient to make the comparison complete. It is necessary to know to what extent the values of the variable tend to concentrate about the average, or, putting it in other words, to what degree they vary or scatter

from the average. This deviation or "scatteration" of the items is called *dispersion*. The closer the items are to the average the less the scatteration, the less the dispersion. The more scattered away from the average the items the greater the dispersion. It is desired, of course, to get a numerical measure of dispersion.

2. Range as a Measure.—Just as there are various kinds of type form or average, there are also various measures of dispersion. The first measure that may suggest itself is the range or difference between the largest and the smallest items. Consider the four groups of values given in Table XXXVII.

TABLE XXXVII.—GROUPS OF VALUES OF FOUR VARIABLES

	I	d_a	II	d_a	III	d_a	IV	d_a
	68	−1	60	− 9	612	− 9	612	− 9
	68	−1	62	− 7	615	− 6	614	− 7
	69	0	65	− 4	619	− 2	617	− 4
	69	0	69	0	621	0	621	0
	69	0	73	4	623	2	625	4
	70	1	76	7	627	6	628	7
	70	1	78	9	630	9	630	9
Totals..........	483	4	483	40	4,347	34	4,347	40
A. M..........	69	0.57	69	5.71	621	4.86	621	5.71

Certain constants for each group are tabulated in Table XXXVIII.

TABLE XXXVIII.—CONSTANTS OF THE FOUR GROUPS

Group	Range	Average, a	Range ÷ Average	δ	$\delta \div a$
I	2	69	$2/69$	0.57	0.008
II	18	69	$18/69$	5.71	0.083
III	18	621	$2/69$	4.86	0.008−
IV	18	621	$2/69$	5.71	0.009

In Group I the range is 2. In group II it is 18. It seems quite evident, then, that there is much greater scatteration, dispersion,

in group II than in group I. Groups II and III have the same range. In group II, however, the range of 18 is with an average of 69, while in group III the range of 18 is with an average of 621. Manifestly, a difference of 18 among things which average 621 is much less noticeable than a difference of 18 among things which average only 69. If the difference in height between the tallest man and the shortest man in a military company is 2 in., it would be said that the height of men in that company is very uniform. There is hardly any variation in height. If it is found that the difference between the longest nose and the shortest nose of the men in the company is 2 in., it would be said that there is very great variation in length of noses.

3. Relative Values.—Unless the variables compared have about the same average, the range must be compared with the average in each case to give a fair idea of the degree of variation. The fourth column of Table XXXVIII gives the ratio of range to average for each of the four groups. In groups I and II, having the same average, these ratios will have the same ratio to each other as did the ranges. One is nine times the other. In groups I and III the ratios of range to average are the same. A range of 2 in things which average 69 is as much, relatively, as a range of 18 in things which average 621. In groups III and IV, having the same range and the same average, the ratios of range to average are equal.

The range has been used here as a *measure of dispersion.* The ratio of range to average would be called a *coefficient of dispersion.* Groups II, III, and IV have the same measure of dispersion, 18, which is nine times that of group I. Groups I, III, and IV have the same coefficient of dispersion, $2/69$, which is one-ninth that of group II.

4. Range May Not Be a Good Measure.—A little consideration will reveal the fact that range is not a very good measure of dispersion. Suppose that in a military company the tallest man is 70 in. tall and the shortest 68 in. Let these two men be replaced by men 71 in. and 67 in. high, respectively. The range is at once doubled, changing from 2 to 4 in. The average height has not been changed. The ratio of range to average has been doubled. The dispersion certainly has been altered but little, though both the measure of dispersion and the coefficient have been doubled. If the heaviest man in a town weighs 160 lb. and the lightest weighs 140 lb., the range in weight is 20 lb. If, now, a fat man

weighing 260 lb. moves into the town, the range is suddenly increased to 120 lb., six times what it was before. The average weight would be increased but little. Taking the town as a whole, the variation in weights has changed but little. In a sample of leaves from a tree, the difference between the lengths of the longest and shortest leaves may be 32 mm., as in Table XX. It is not unlikely that there are a few stray leaves on the tree shorter than any in the sample or a few longer than any in the sample. A second sample might have a much greater range than the first, and yet both samples be good representations of the entire universe from which they were selected. A measure so seriously affected by what may be called accidental variations can hardly be considered a good measure.

5. Deviations from the Average.—To get the coefficient of dispersion the range was compared with the average. The deviation of the extreme item from the average might have suggested itself as a measure of dispersion. It is at once seen that there is the same objection that there is to the range. If, however, the deviation of each time from the average can be considered in the measure of dispersion, a much more satisfactory measure will result. This may be illustrated by the items in groups III and IV of Table XXXVII. In group III it is seen that the next item larger and the next smaller than the average, 623 and 619, differ from the average a less amount than do the corresponding items, 625 and 617, in group IV. The same is true for the next items beyond these. Though the largest and smallest items in each group differ from the average by the same amount, yet it is evident that in the group *as a whole*, the items of group III cluster around the average more closely than do those of group IV. The dispersion in group III is then less than that in group IV. The scatteration is greater in group IV than in group III.

6. Average Deviation.—The *average deviation* from the average of the group at once suggests itself as a measure of dispersion. The closer the items cluster about the average, the smaller will be the average deviation. The more they are scattered away from the average, the greater will be the average deviation. In Table XXXVII the deviations from the average are tabulated for each group. The columns are headed d_a. In summing the deviations, their absolute values are used, since positive and negative deviations of the same size would have the same influ-

ence on the degree of scatteration. The average absolute deviations δ are tabulated in Table XXXVIII.

$$\delta = \frac{\Sigma|d_a|}{N}.$$

7. Coefficient.—Groups II and IV have the same average deviation of $\delta = 5.71$. *But an average deviation of 5.71 among* things whose average size is 69 is much greater than an average deviation of 5.71 among things whose average size is 621. So again, when the average sizes differ materially, it is necessary to compare average deviation with average size of item to get a coefficient of dispersion.

Coefficient of dispersion $= \delta/a$.

8. The Average Used.—The average used from which to take deviations may be either arithmetic mean, median, or mode. Whichever is used, the corresponding subscript may be attached to δ. To get the coefficient of dispersion, δ is divided by that average from which δ was derived.

$$\delta_a = \frac{\Sigma|X - a|}{N} = \frac{\Sigma|d_a|}{N}.$$

$$\delta_M = \frac{\Sigma|X - M|}{N} = \frac{\Sigma|d_M|}{N}.$$

$$\delta_Z = \frac{\Sigma|X - Z|}{N} = \frac{\Sigma|d_Z|}{N}.$$

The corresponding coefficients are

$$\frac{\delta_a}{a}, \frac{\delta_M}{M}, \text{ and } \frac{\delta_Z}{Z}.$$

9. Examples from Frequency Tables.—Working from a frequency table, let it be required to find δ_a and δ_M for freshman weights from Table XIII.

TABLE XXXIX.—AVERAGE DEVIATION FROM ARITHMETIC MEAN OF FRESHMAN WEIGHTS

Size of item, in pounds X	Frequency f	fX	d_a	fd_a
105	15	1,575	−37.25	− 558.75
116	43	4,988	−26.25	−1,128.75
127	138	17,526	−15.25	−2,104.50
138	162	22,356	− 4.25	− 688.50
149	129	19,221	6.75	870.75
160	82	13,120	17.75	1,455.50
171	35	5,985	28.75	1,006.25
182	16	2,912	39.75	636.00
193	5	965	50.75	253.75
204	3	612	61.75	185.25
215	1	215	72.75	72.75
Total	629	89,475	......	8,960.75

$$a = \frac{89,475}{629} = 142.25. \qquad \delta_a = \frac{8,960.75}{629} = 14.24.$$

$$\frac{\delta_a}{a} = \frac{14.24}{142.25} = 0.1001.$$

TABLE XL.—AVERAGE DEVIATION OF FRESHMAN WEIGHTS FROM THE MEDIAN

Size of item, in pounds X	Frequency f	d_M	fd_M
105	15	−35.55	− 533.25
116	43	−24.55	−1,055.65
127	138	−13.55	−1,869.90
138	162	− 2.55	− 413.10
149	129	8.45	1,090.05
160	82	19.45	1,594.90
171	35	30.45	1,065.75
182	16	41.45	663.20
193	5	52.45	262.25
204	3	63.45	190.35
215	1	74.45	74.45
Total	629	......	8,812.85

$$M = 140.55 \text{ (see p. 116, Chap. IV).}$$

$$\delta_M = \frac{8,812.85}{629} = 14.01.$$

$$\frac{\delta_M}{M} = \frac{14.01}{140.55} = 0.099.$$

δ_M must be less than δ_a, since, as has been proved, the sum of the absolute values of the deviations from the median is a minimum. The median is, for this reason, more logical to use for average deviation than one of the other averages.

10. Standard Deviation.—In order to avoid the difficulty of negative deviations, each deviation may be squared. The squares of the deviations are then added, the sum divided by the number of observations, and the square root of the quotient taken. This result is root-mean-square of the deviations. If there are N observations, and deviations are taken from some number X', the root-mean-square deviation is

$$s = \sqrt{\frac{\Sigma d_{X'}^2}{N}},$$

in which $d_{X'}$ is deviation from X'. Root-mean-square deviation from the arithmetic mean is known as *standard deviation*. Its mathematical properties make it of great importance in the study of statistical distributions and the theory of errors. The standard deviation is

$$\sigma = \sqrt{\frac{\Sigma d_a^2}{N}},$$

where σ is read *sigma* or *standard deviation.*

11. Relation between s and σ.—The following relation always holds between s and σ,

$$s^2 = \sigma^2 + \bar{x}'^2,$$

where

$$\bar{x}' = a - X' = \frac{\Sigma d_{X'}}{N}.$$

Proof:

$$s^2 = \frac{\Sigma(X - X')^2}{N},$$

$$\bar{x}'^2 = (a - X')^2,$$

$$s^2 - \bar{x}'^2 = \frac{\Sigma(X - X')^2 - N(a - X')^2}{N}$$

$$= \frac{\Sigma X^2 - 2X'\Sigma X + NX'^2 - Na^2 + 2NaX' - NX'^2}{N}$$

$$= \frac{\Sigma X^2 - Na^2}{N} \qquad \text{Since } Na = \Sigma X.$$

$$= \frac{\Sigma(d_a + a)^2 - Na^2}{N} \qquad \text{Since } d_a = X - a.$$

$$= \frac{\Sigma d_a^2 + 2a\Sigma d_a + Na^2 - Na^2}{N}$$

$$= \frac{\Sigma d_a^2}{N} \qquad \text{Since } \Sigma d_a = 0.$$

$$= \sigma^2. \qquad \text{By definition.}$$

$$\therefore \quad s^2 = \sigma^2 + \bar{x}'^2.$$

Multiplying both sides of

$$\sigma^2 = s^2 - \bar{x}'^2$$

by N gives

$$N\sigma^2 = Ns^2 - N\bar{x}'^2.$$

Ns^2 is the sum of the squares of deviations from any arbitrary number X'. $N\sigma^2$ is the sum of the squares of deviations from the arithmetic mean. Therefore, from the last equation above, it is true that:

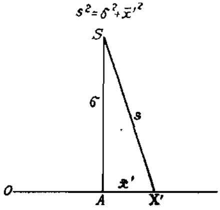


FIG. 42.

The sum of the squares of deviations from the arithmetic mean is a minimum. So the arithmetic mean is the logical average from which to measure dispersion when the squares of deviations are used.

12. Relation Graphically.—Knowing the values of σ and $\bar{x}'$, the value of s may at once be obtained graphically as follows: On the x-axis at the value of the arithmetic mean, A (Fig. 42), erect a perpendicular, AS, equal to σ measured to the scale of the x-axis. On the x-axis lay $\bar{x}'$ from A to the point X'. Draw $X'S$. Since

$$(X'S)^2 = (AS)^2 + (AX')^2,$$
$$(X'S)^2 = \sigma^2 + \bar{x}'^2.$$

Therefore,

$$X'S = s.$$

Conversely if s and $\bar{x}'$ are known, σ can at once be determined graphically.

13. Other Names.—Standard deviation, σ, is also known as *dispersion, index of variability, mean error, root-mean-square-error*, etc. In this constant it is always understood that deviations are taken from the arithmetic mean.

14. A Measure of Consistency.—Since the deviations are squared, large deviations have more influence in determining σ than in average deviation. This is, to some extent, offset by taking the square root of the average square of deviations. In measuring physical quantities, the presence of many large deviations from the arithmetic mean would reveal inconsistency in the observed measurements. So σ is used in measuring consistency of the set of observations. Suppose that the surveyor, in taking a number of measurements of the length of a line by a yardstick, finds that the standard deviation is larger than when a steel tape is used. He at once concludes that the tape measurements are more consistent than the yardstick measurements. The same conclusions may be drawn in regard to any variable that approximately follows a normal distribution.

15. The mechanical interpretation of standard deviation is that it is the radius of gyration of the area under the frequency curve about the ordinate through the center of gravity of that area. This can be readily seen by those familar with the meaning of *radius of gyration*.

The radius of gyration of a rectangular histogram about an axis through its center of gravity is the square root of the sum of the products of the area of each rectangle multiplied by the square of the distance of its center from the axis through the center of gravity, divided by the total area. The width of each rectangle being equal to one class interval may be regarded as unity. The areas of the rectangles are then equal to class frequencies. The ordinate at the arithmetic mean passes through the center of gravity of the entire area and so may be used as the axis of gyration. The distances of centers of rectangles from the axis are deviations of the midpoints of classes from the arithmetic mean. Then the product of the area of a rectangle by the square of the distance from the axis is the product of frequency by square of deviation from arithmetic mean. The sum of these products is

the sum of the squares of deviations from the arithmetic mean for every item in the set. The entire area is total frequency. So the quotient of the sum of the products of the area of each rectangle times the square of its distance from the axis, divided by the total area, equals the average square deviation from the arithmetic mean. Or, in other words, the square of the radius of gyration equals σ^2. Now let the width of rectangles decrease indefinitely, their number increasing without limit, so that the rectangular histogram approaches the frequency curve as a limit. Then the radius of gyration of the rectangular histogram approaches the radius of gyration of the area under the frequency curve as a limit, and the standard deviation of the sample shown in the rectangular histogram approaches as a limit the standard deviation of the universe from which samples were taken if they are truly representative.

16. Meaning of Radius of Gyration.—The meaning of radius of gyration is that if a mass is rotating about an axis, the radius of gyration is such a distance that if the entire mass were concentrated at that distance from the axis the rotational effect would be the same as for the actual distributed mass. In other words, it may be called an average rotational radius for all the particles of the rotating mass. In the same way standard deviation is an average deviation, or a measure of dispersion.

17. Relation between s and σ.—It is a well-known theorem in mechanics that the square of the radius of gyration of an area about an axis not through the center of gravity equals the square of the radius of gyration about a parallel axis through the center of gravity plus the square of the distance between the axes. From this follows at once the relationship:

$$s^2 = \sigma^2 + \bar{x}'^2.$$

There are a number of interesting analogies between the theorems of mechanics and those of statistical method.

18. Examples.—The method of computing standard deviation from such sets of values of variables as those given in Table XXXVII is obvious. A column of d_a^2 would be added to the table. This column would be totaled, giving Σd_a^2. This total would be divided by the number of items, N, and the square root of the quotient found. For example, for group I,

$$\Sigma d_a^2 = 4, N = 7.$$
$$\sigma = \sqrt{4/7} = 0.76.$$

The average deviation,

$$\delta = 0.57.$$

For groups II and IV,

$$\Sigma d_a^2 = 292.$$
$$\sigma = \sqrt{292/7} = 6.5.$$

The average deviation,

$$\delta = 5.7.$$

For group III,

$$\Sigma d_a^2 = 242.$$
$$\sigma = \sqrt{242/7} = 5.9.$$

The average deviation,

$$\delta = 4.86.$$

As would be expected, in each case

$$\sigma > \delta.$$

19. Coefficient of Dispersion.—As in average deviation, here it is not fair to compare σ of either group I or II with σ of either group III or IV, on account of the great difference in averages. So here, as was done before, coefficients of variation should be used for comparison if the averages are not about equal. The *coefficient of dispersion*, or *coefficient of variation*, equals σ/a. For the different groups of Table XXXVII, the coefficients of variation are

Group	$\frac{\sigma}{a}$
I	0.01
II	0.10
III	0.009
IV	0.01

The coefficient is sometimes multiplied by 100 and is thus stated as a per cent.

The relative size of standard deviation and the average from which it is derived is important. If the standard deviation of the heights of the men in the military company of Sec. 2 of this chapter is 1 in. and that of the lengths of their noses is also 1 in., the spread, or scatter, of heights is the same as that of nose length. But if the average height of the men is 69 in., the standard deviation is insignificantly small. If the average nose length is 1.75

in., a standard deviation of 1 in. is very significantly large. The standard deviation of weights of a random sample of elephants may be more than the standard deviation of weights of a random sample of cats and yet be insignificantly small as compared with that of cats. So, for purposes of comparison, relative values must be used.

20. Computation from Frequency Table, Direct Method.—The method of working from a frequency table may be shown by use of Table XIV of freshmen heights. The table for computing σ by direct method of deviations from the arithmetic mean is shown as Table XLI.

TABLE XLI.—STANDARD DEVIATION OF HEIGHTS, DIRECT METHOD

X	f	fX	d_a	d_a^2	fd_a^2
60.5	3	181.5	−7.68	58.9824	176.95
62.0	8	496.0	−6.18	38.1924	305.54
63.5	33	2,095.5	−4.68	21.9024	722.78
65.0	51	3,315.0	−3.18	10.1124	515.73
66.5	115	7,647.5	−1.68	2.8224	324.58
68.0	156	10,608.0	−0.18	0.0324	5.05
69.5	148	10,286.0	1.32	1.7424	257.88
71.0	64	4,544.0	2.82	7.9524	508.95
72.5	43	3,117.5	4.32	18.6624	802.48
74.0	5	370.0	5.82	33.8724	169.36
75.5	3	226.5	7.32	53.5824	160.75
Totals..............	629	42,887.5	......	....	3,950.05

$$a = \frac{42{,}887.5}{629} = 68.18. \qquad \sigma = \sqrt{\frac{3{,}950.05}{629}} = \sqrt{6.2799} = 2.51.$$

$$\frac{\sigma}{a} = \frac{2.51}{68.18} = 0.037.$$

The first column, size of item X, shows midpoints of classes. The second column is frequency f. The third column is fX. The total of this column divided by N gives a, the arithmetic mean.

$$a = \frac{42{,}887.5}{629} = 68.18.$$

The fourth column gives d_a, deviations from the arithmetic mean. The fifth column gives the squares of the deviations. If

this column were added, the sum would give only one d_a^2 for each class. The sum of d_a^2 must be obtained for all N items; so the sixth column of fd_a^2 is formed. The total of this column is Σd_a^2 for the entire table. Then

$$\sigma = \sqrt{\frac{\Sigma d_a^2}{N}},$$

$$\sigma = \sqrt{\frac{3{,}950.05}{629}},$$

$$\sigma = 2.51.$$

The coefficient of dispersion is

$$\frac{\sigma}{a} = \frac{2.51}{68.18}$$

$$= 0.037.$$

TABLE XLII.—SHORT-CUT METHOD

X	f	$d_{X'}$	$fd_{X'}$	$fd^2_{X'}$
60.5	3	−7.5	− 22.5	168.75
62.0	8	−6.0	− 48.0	288.00
63.5	33	−4.5	−148.5	668.25
65.0	51	−3.0	−153.0	459.00
66.5	115	−1.5	−172.5	258.75
68.0	156	0	0	0
69.5	148	1.5	222.0	333.00
71.0	64	3.0	192.0	576.00
72.5	43	4.5	193.5	870.75
74.0	5	6.0	30.0	180.00
75.5	3	7.5	22.5	168.75
Totals.........	629	.....	+115.5	3,971.25

$$X' = 68.0.$$

$$\bar{x}' = \frac{+115.5}{629} = 0.18.$$

$$a = 68.0 + 0.18 = 68.18.$$

$$\sigma = \sqrt{\frac{3{,}971.25}{629} - (0.18)^2}$$

$$= \sqrt{6.3136 - 0.0324}$$

$$= \sqrt{6.2812}$$

$$= 2.51.$$

21. Short-cut.—Table XLII shows the method of computation by use of deviations from some arbitrary number X'. This is the short-cut method. X' should be chosen as one of the mid-points of classes; so the deviations for that class will be 0. Take that class nearest the middle of the table or nearest the estimated arithmetic mean or the class of greatest frequency. In this table the sixth class satisfies all these conditions.

$$X' = 68.0.$$

The first and second columns are the same as in Table XLI. The third column, $d_{X'}$, is deviation from X'. The deviation for $X = 68.0$ is 0. The deviations above and below this are multiples of the class interval, positive for $X > 68.0$ and negative for $X < 68.0$. Each deviation is multiplied by the corresponding frequency to give the fourth column. The total of this column, $\Sigma d_{X'}$, divided by N gives $\bar{x}'$, the deviation of the arithmetic mean a, from X', the arbitrary origin of measurements. The following equation has previously been stated:

$$a = X' + \frac{\Sigma d_{X'}}{N},$$

or

$$\frac{\Sigma d_{X'}}{N} = a - X'.$$

$\bar{x}'$ was defined on page 137 as $a - X'$,

$$\therefore \quad \bar{x}' = \frac{\Sigma d_{X'}}{N},$$

Here

$$\bar{x}' = \frac{115.5}{629} = 0.18\cdot$$

$$\begin{aligned} a &= X' + \bar{x}' \\ &= 68.00 + 0.18 \\ &= 68.18. \end{aligned}$$

The entries in the fifth column are formed by taking the product of corresponding entries in the third and fourth columns, giving $fd^2_{X'}$. The total of this column divided by N is s^2. From page 137,

$$s = \sqrt{\frac{\Sigma d^2_{X'}}{N}}.$$

Here,

$$s^2 = \frac{3,971.25}{629}$$

From

$$s^2 = \sigma^2 + \bar{x}'^2$$
$$\sigma^2 = s^2 - \bar{x}'^2.$$

From this table, then,

$$\sigma = \sqrt{\frac{3{,}971.25}{629} - (0.18)^2}$$
$$= \sqrt{6.3136 - 0.0324}$$
$$= \sqrt{6.2812}$$
$$= 2.51.$$

The coefficient of dispersion is

$$\frac{\sigma}{a} = \frac{2.51}{68.18}$$
$$= 0.037.$$

In general,

$$\sigma = \sqrt{s^2 - \bar{x}'^2}$$
$$= \sqrt{\frac{\Sigma d^2_{X'}}{N} - \left(\frac{\Sigma d_{X'}}{N}\right)^2}$$
$$= \sqrt{\frac{\Sigma(X - X')^2}{N} - (a - X')^2}$$
$$= \sqrt{\frac{\Sigma(X - X')^2 - N(a - X')^2}{N}}.$$

Whichever form is preferred may be used.

22. If $X' = 0$.—It may be of interest to see what form the formula for σ takes when $X' = 0$. Deviations of the items from 0 are the values of the items themselves.

So

$$d_{X'} = X,$$
$$\bar{x}' = a - X' = a,$$
$$s = \sqrt{\frac{\Sigma d^2_{X'}}{N}}$$
$$= \sqrt{\frac{\Sigma X^2}{N}}.$$
$$\sigma = \sqrt{s^2 - \bar{x}'^2}$$
$$= \sqrt{\frac{\Sigma X^2}{N} - a^2}$$
$$= \sqrt{\frac{\Sigma X^2 - Na^2}{N}}.$$

TABLE XLIII.—STANDARD DEVIATION OF FRESHMAN HEIGHTS FROM $X' = 0$

X	f	fX	fX
60.5	3	181.5	10,980.75
62.0	8	496.0	30,752.00
63.5	33	2,095.5	133,064.25
65.0	51	3,315.0	215,475.00
66.5	115	7,647.5	508,558.75
68.0	156	10,608.0	721,344.00
69.5	148	10,286.0	714,877.00
71.0	64	4,544.0	322,624.00
72.5	43	3,117.5	226,018.75
74.0	5	370.0	27,380.00
75.5	3	226.5	17,100.75
Totals............	629	42,887.5	2,928,175.25

$$a = \frac{42{,}887.5}{629} = 68.18$$

$$\begin{aligned}\sigma &= \sqrt{\frac{2{,}928{,}175.25}{629} - (68.18)^2} \\ &= \sqrt{4{,}655.28 - 4{,}648.51} \\ &= \sqrt{6.77} \\ &= 2.51.\end{aligned}$$

Table XLIII shows the computation for freshman heights. The formula is very simple, but the work of calculation is increased over that of Table XLII.

The formula

$$\sigma = \sqrt{\frac{\Sigma X^2 - Na^2}{N}}$$

may be written

$$\sigma^2 = \frac{\Sigma X^2}{N} - a^2.$$

This translated into English reads: *The square of standard deviation equals the average square of the items diminished by the square of the average item.*

If distributions are approximately normal, it has been found empirically that

$$\frac{\delta}{\sigma} = \frac{4}{5} = 0.8$$

nearly.

If the distribution is that of the normal probability curve, the ratio δ/σ can be proved to be 0.7979.[1]

23. Labor of Computation.—The labor of computation of average deviation is a little less than that for standard deviation. If computing machines or product tables and tables of squares are available, the amount of labor is about the same for each. Where the measure of dispersion is to be used in further mathematical work, the standard of deviation lends itself to algebraic processes much more readily than does average deviation.

24. σ is an Average.—Standard deviation is an average in the sense of the general definition of average. If

$$d_1 = d_2 = d_3 = \ . \ . \ . \ = d_N = k,$$

then

$$\sqrt{\frac{\Sigma d^2}{N}} = \sqrt{\frac{Nk^2}{N}} = k$$

equals each of the d's.

25. $\sigma = \sqrt{npq}$.—In Chap. II, page 31, Theorem III, it was stated that if p is the probability of success of a given event in a single trial and q is the probability of its failure, then the probability of r successes and $n - r$ failures in n trials is

$${}_nC_r p^r q^{n-r},$$

the $(n - r + 1)$th term in the binomial expansion of

$$(p + q)^n.$$

Always

$$p + q = 1.$$

The theoretical number of times, then, that one would expect to get r tails and $n - r$ heads in tossing n coins N times is $N{}_nC_r p^r q^{n-r}$. In other words, the most probable frequencies of occurrence of the different possible combinations of heads and tails in tossing n coins N times are equal to the terms of the binomial expansion of

$$N(p + q)^n,$$

where p is probability of throwing tails in a single toss of one coin and q is the probability of throwing heads. The various probabilities for seven dimes were worked out on page 31, Chap. II, and the actual frequencies in an experiment of tossing the seven dimes 500 times were compared with the theoretical frequencies in Table XVII. On page 78

$$\sigma = \sqrt{npq}$$

was used as a constant.

[1] JONES, "First Course in Statistics," p. 246.

It is now desired to show that this σ is standard deviation of the entire distribution of most probable frequencies as given by the terms of

$$N(p + q)^n.$$

Since N is a common factor of every term, the terms of the expansion of

$$(p + q)^n$$

are the fractions of N giving the corresponding frequencies. Or, if N is regarded as unity, these fractions themselves may be called the frequencies. Now form a frequency table with columns for getting arithmetic mean and standard deviation, as in Table XLIV. The origin of measurements for deviations X' is taken at 0, as was done in Table XLIII.

The entries in f column are the terms of the expansion of $(p + q)^n$.

Therefore,

$$N = \Sigma f = 1.$$

Since

$$(p + q) = 1,$$
$$(p + q)^n = 1.$$

In the fX column, nq is a factor of every term.

$$\begin{aligned}\Sigma fX &= nq\left\{p^{n-1} + (n-1)p^{n-2}q + \frac{(n-1)(n-2)}{1 \cdot 2}p^{n-3}q^2 \right. \\ &\qquad \left. + \frac{(n-1)(n-2)(n-3)}{1 \cdot 2 \cdot 3}p^{n-4}q^3 + \ . \ . \ . + q^{n-1}\right\} \\ &= nq(p+q)^{n-1} \\ &= nq.\end{aligned}$$

since

$$(p + q)^{n-1} = 1$$

In the fX^2 column, nq is a common factor of every term.

$$\begin{aligned}\Sigma fX^2 &= nq\left\{p^{n-1} + 2(n-1)p^{n-2}q + 3\,\frac{(n-1)(n-2)}{1 \cdot 2}\,p^{n-3}q^2 \right. \\ &\qquad \left. + \frac{4(n-1)(n-2)(n-3)}{1 \cdot 2 \cdot 3}\,p^{n-4}q^3 + \ . \ . \ . + nq^{n-1}\right\} \\ &= nq\left\{[p^{n-1} + (n-1)p^{n-2}q + \frac{(n-1)(n-2)}{1 \cdot 2}\,p^{n-3}q^2 \right. \\ &\quad + \frac{(n-1)(n-2)(n-3)}{1 \cdot 2 \cdot 3}\,p^{n-4}q^3 + \ . \ . \ . + q^{n-1}] \\ &\quad + [(n-1)p^{n-2}q + 2\,\frac{(n-1)(n-2)}{1 \cdot 2}\,p^{n-3}q^2\end{aligned}$$

TABLE XLIV.—ARITHMETIC MEAN AND STANDARD DEVIATION FOR DISTRIBUTIONS PROPORTIONAL TO THE TERMS OF $(p + q)^n$

Number of heads, X	Frequency f	fX	fX^2
0	p^n	0	0
1	$np^{n-1}q$	$np^{n-1}q$	$np^{n-1}q$
2	$\frac{n(n-1)}{1 \cdot 2} p^{n-2}q^2$	$n(n-1)p^{n-2}q^2$	$2n(n-1)p^{n-2}q^2$
3	$\frac{n(n-1)(n-2)}{1 \cdot 2 \cdot 3} p^{n-3}q^3$	$\frac{n(n-1)(n-2)}{1 \cdot 2} p^{n-3}q^3$	$\frac{3n(n-1)(n-2)}{1 \cdot 2} p^{n-3}q^3$
4	$\frac{n(n-1)(n-2)(n-3)}{1 \cdot 2 \cdot 3 \cdot 4} p^{n-4}q^4$	$\frac{n(n-1)(n-2)(n-3)}{1 \cdot 2 \cdot 3} p^{n-4}q^4$	$\frac{4n(n-1)(n-2)(n-3)}{1 \cdot 2 \cdot 3} p^{n-4}q^4$
.	.	.	.
.	.	.	.
.	.	.	.
n	q^n	nq^n	n^2q^n
Totals.......	1	nq	$nq + n^2q^2 + nq^2$

$$+ \frac{3(n-1)(n-2)(n-3)}{1 \cdot 2 \cdot 3} p^{n-4}q^3 + \ . \ . \ . + (n-1)q^{n-1}] \Big\}$$

$$= nq \Big\{ (p+q)^{n-1} + (n-1)q[p^{n-2} + (n-2)p^{n-3}q + \frac{(n-2)(n-3)}{1 \cdot 2} p^{n-4}q^2 + \ . \ . \ . + q^{n-2}] \Big\}$$

$$= nq\{(p+q)^{n-1} + (n-1)q(p+q)^{n-2}\}$$

$$= nq\{1 + (n-1)q\}$$

$$= nq + n^2q^2 - nq^2.$$

$$a = \frac{\Sigma X}{N} = \frac{nq}{1} = nq.$$

With

$$X' = 0 \text{ (see p. 145)},$$

$$\bar{x}' = a,$$

$$\sigma^2 = \frac{\Sigma f X^2}{N} - \bar{x}'^2$$

$$= \frac{nq + n^2q^2 - nq^2}{1} - n^2q^2$$

$$= nq - nq^2$$

$$= nq(1-q)$$

$$= npq.$$

So

$$\sigma = \sqrt{npq}.$$

If p and q are the probabilities of throwing tails or heads, respectively, of a single coin in a single trial, then

$$p = q = \tfrac{1}{2}.$$

If p is the probability of throwing a 5 in a single throw of a single die, then

$$p = \tfrac{1}{6}.$$

The probability of failure to get a 2 is

$$q = \tfrac{5}{6}.$$

Other values of p and q apply to other events. The above value of standard deviation applies to distributions which follow the terms of the binomial expansion of

$$(p + q)^n,$$

no matter what values p and q may have.

26. Quartiles and Other Division Points.—Sometimes it is desirable to regard the distribution of the middle portion of the array only. This, of course, leaves the extreme items entirely out of

consideration. This would be the case if there is reason for believing that the extreme items are merely freaks that should not influence the conclusions in regard to distribution. Moreover, the statistical constant called probable error, discussed later, is concerned with the range of the middle half of the array. So it becomes at once a matter of importance to find the two division points of the array known as quartiles. Definitions and formulas for finding these and other division points in general, including the median, will now be given.

27. Placing in Array.—Given N items. Place them in array on a horizontal base line, letting each item occupy unit space on this line. The length of this line is N. In arraying the items, place the smallest item at the left-hand end of the base line, occupying the segment from 0 to 1, the next larger item occupying the segment from 1 to 2, etc. The largest item will occupy the Nth unit segment of the base line. Now draw a curve through the tops of the items as arrayed.

28. The median is the value of the ordinate to this curve half way through the array. This is the ordinate at the midpoint of the base line. It stands at the horizontal scale mark, $N/2$. It is evident, if N is even, that exactly half the items are each less than the median and half the items are each greater than the median. If N is odd, the median ordinate will stand at the middle of the center unit segment of the base line. The item standing on this center segment may be regarded as belonging half to those items less than the median and half to those items greater than the median. It is true that exactly half the items lie to the left of the median and half to the right, the middle item being divided into halves by the median. It would, then, be correct to say that half the items are less than the median and half are larger than the median. In either case it is strictly true that there are as many items less than the median as there are greater than the median.

29. The first quartile is the value of the ordinate one-fourth *of the way through the array from the zero end of the base line.* It is the ordinate at one-fourth of the length of the base line from its zero end or three-fourths of the length of the base line from the other end. It stands at the horizontal scale mark $N/4$. It is evident if N is a multiple of 4, of the form $N = 4n$, that exactly one-fourth of the items are each less than the first quartile and three-fourths are each greater than the first quartile. If N is of

the form $4n + 1$, $4n + 2$, or $4n + 3$, then $N/4$ equals $n + \frac{1}{4}$, $n + \frac{2}{4}$, or $n + \frac{3}{4}$, respectively. The ordinate measuring the first quartile is then at $\frac{1}{4}$, $\frac{2}{4}$, or $\frac{3}{4}$ of a unit from the beginning of the unit segment upon which it stands. If each item is regarded as occupying the entire unit segment upon which it stands, then $\frac{1}{4}$, $\frac{2}{4}$, or $\frac{3}{4}$, respectively, of the item on that unit segment is to the left of the quartile and $\frac{3}{4}$, $\frac{2}{4}$, or $\frac{1}{4}$ of it to the right of the quartile. It is true, then, that exactly $\frac{1}{4}$ of the N items are to the left of the first quartile and $\frac{3}{4}$ of them to the right. If one may regard $\frac{1}{4}$, $\frac{2}{4}$, or $\frac{3}{4}$, respectively, of the item standing on this segment as belonging with those items less than the first quartile, and $\frac{3}{4}$, $\frac{2}{4}$, or $\frac{1}{4}$, respectively, of this item as belonging with those items greater than the first quartile, then it would be correct to say that 25 per cent of the items are each less than the first quartile and 75 per cent are each greater than the first quartile.

To illustrate: If $N = 24$, $N/4 = 6$, and the ordinate measuring the first quartile is at six units from the beginning of the base line; 6 items, 25 per cent of the total are each less than the first quartile and 18 items, 75 per cent of the total, are each greater than the first quartile. If $N = 25$, $N/4 = 6.25$. The ordinate measuring the first quartile is at 6.25 units from the zero end of the base line. Thus $6\frac{1}{4}$ items are regarded as each less than the first quartile, and $18\frac{3}{4}$ items as each greater than the first quartile. Again 25 per cent of the items are each less than the first quartile and 75 per cent of them each greater.

From the above definition and illustration, the definition of the *first quartile* may be justified as being such a quantity that 25 per cent of the items are each less than the first quartile and 75 per cent are each larger. The first quartile need not be equal to one of the items in the array. If it happens to equal one of the items, that item is divided and regarded as belonging half to those items less than the first quartile and half to those items greater.

30. The third quartile is the value of the ordinate to the curve three-fourths of the way through the array from the zero end of the base line. It is the ordinate at three-fourths of the length of the base line from the zero end and one-fourth of the length of the base line from the other end. It stands at the horizontal scale mark, $3N/4$. If N is a multiple of 4, it is evident that exactly three-fourths of the items are each less than the third quartile and one-fourth are each greater. If N is not a multiple of 4, assumptions are made similar to those used in the discussion

of the first quartile. Under these assumptions the *third quartile* may be defined as being such a quantity that 75 per cent of the items are each less than the third quartile and 25 per cent are each larger.

31. Other Division Points.—Sometimes octiles, deciles, or percentiles are used. The *first octile* is the value of the ordinate to the curve at one-eighth of the way through the array, the *second octile*, the value of the ordinate two-eighths of the way through the array, and so on for the other octiles. The *first decile* is the value of the ordinate to the curve at one-tenth of the way through the array, the *second decile* two-tenths, and so on for each decile. Each of the above division points may be expressed as *a percentile*, thus making percentile a more general term for a division point. The pth *percentile* is the ordinate to the curve at p per cent of the way through the array.

32. Notation.—The following notation will be used:

M = median.
Q_1 = first quartile.
Q_3 = third quartile.
O_1 = first octile.
O_3 = third octile.
D_4 = fourth decile.
P_p = pth percentile.
$P_{50} = D_5 = O_4 = Q_2 = M.$

Similar equalities in different parts of the array are readily seen to hold.

33. Significance of Division Points.—As was shown in the chapter on Averages, M is that size such that the sum of the absolute values of the deviations of all items from it is less than the sum of the absolute values of the deviations from any other size. Q_1 and Q_3 are such ordinates that exactly the middle half of the items of the array lie between them. $Q_3 - Q_1$ is called the *interquartile range*. It is an important quantity. Dividing the array by means of any two octiles or two deciles segregates a definite part of the array which one may wish to study. The percentiles divide the array into 100 parts each containing 1 per cent of the total number of items in the array. By means of proper selection of percentiles, the array can be divided in almost any manner for purposes of study and comparison. If population of a locality is arrayed according to wealth, that portion of population between P_{50} and P_{95} is sometimes called the *middle*

class. It excludes the poorer half and the richest 5 per cent of the population. Not infrequently this middle class has about half of the wealth.

34. Determination of Median, Quartiles, and Other Division Points.—If there is a relatively small number N of items in a sample, it is easy to determine the division points desired by placing the items in array and counting up to the point in question.

If N is relatively large, a frequency table may either be given or be formed from the original tabulation of measurements. Suppose an array of items is given as shown by a frequency table consisting of the first two columns of Table XLV. Although N $(= 24)$ is here relatively small, the table serves the purpose of illustration. Form the cumulative columns F and F'.

TABLE XLV.—DETERMINATION OF DIVISION POINTS

Size X	Frequency f	Cumulation from *lowest class up* F	Cumulation from *highest class down* F'
3–7	2	2	24
8–12	5	7	22
13–17	9	16	17
18–22	4	20	8
23–27	3	23	4
28–32	1	24	1
	N 24		

On a horizontal line (base line) (Fig. 43) plot the class boundaries vertically. Plot the lowest boundary, 2.5, at 0. Plot the next boundary, 7.5, at 2, the frequency of the first class. Plot all the class boundaries so that consecutive boundaries are at distances apart equal to the corresponding class frequencies. This places them at values of the cumulative frequencies F. The ordinates representing the class boundaries of the table are drawn in dotted lines. Connect the top of each dotted line to the top of the next by a straight line. These straight lines form a broken line, ordinates to which measure the size of items. *Let each item occupy a unit segment of the base line, the smallest item standing on the segment 0 to 1, the next larger on 1 to 2, and so on to the*

largest, which stands on the segment 23 to 24. Not knowing the exact distribution of sizes in any class, assume uniform variation throughout each class. The size of any item is then measured by the ordinate to the broken line at the midpoint of the unit segment of the base line on which that item stands. Thus the size of the fifth item of the array (third in the second class) is, under the assumption of uniform variation, equal to 10, the ordinate to the broken line at the midpoint of the base line segment 4 to 5. Ordinates at the proper places measure the required division points.

35. Derivations of Formulas.—Now determine a formula for finding any division point. This includes the formula for median which was previously derived in Chap. IV. The process is repeated here for the sake of generality for all division points.

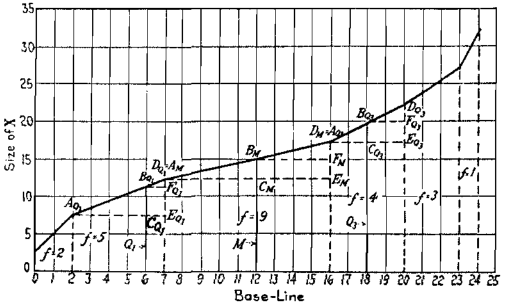


FIG. 43.

N = total number of items,

c = class interval.

f = frequency of class in which the required point lies.

l = lower boundary of class in which the required point lies.

l' = upper boundary of class in which the required point lies.

F_p = cumulative frequency up to the class in which the required point lies. For the median, $P = M$, for first quartile, $P = Q_1$, and similarly for other division points.

F'_P = cumulative frequency down to the class in which the required point lies.

i = number of units of base line that must be passed from the lower boundary of the class in which the required point lies to reach the required point.

i' = number of units that must be passed from the upper boundary to the required point.

x = amount that must be added to l to give the value sought.

x' = amount that must be subtracted from l' to give the value sought.

Suppose it is desired to find the median M. In the given table $N = 24$, $c = 5$. $N/2 = 12$, the base-line scale mark at which the median stands. At 12 erect the ordinate to B_M. This ordinate measures M. Examining the F column of the table, it is found that the twelfth item is in the third class. So $f = 9$, $l = 12.5$, $F_M = 7$, $i = N/2 - F_M = 12 - 7 = 5$. As shown in Fig. 43, $x = C_MB_M$, the amount that must be added to $l = 7\text{-}A_M$ the lower boundary of the class in which M lies, to give the value of M. $c = E_MD_M$, $f = A_ME_M$, $i = A_MC_M$. From the similar triangles $A_MB_MC_M$ and $A_MD_ME_M$

$$\frac{C_MB_M}{E_MD_M} = \frac{A_MC_M}{A_ME_M}.$$

Substituting the above values and solving for x,

$$\frac{x}{c} = \frac{i}{f},$$

$$x = \frac{ci}{f}.$$

Therefore,

$$M = l + x = l + \frac{ci}{f}.$$

Using the values of the given example,

$$M = 12.5 + \frac{5 \times 5}{9},$$

$$M = 12.5 + 2.8,$$

$$M = 15.3.$$

Working from the upper values down: $N = 24$, $N/2 = 12$. From the F' column of the table it is seen that the twelfth item from the upper end of the distribution is in the fourth class. So $f = 9$, $l' = 17.5$, $F'_M = 8$,

$$i' = \frac{N}{2} - F'_M = 12 - 8 = 4.$$

In the figure, $x' = F_M D_M$, the amount to be subtracted from $l' = 16\text{-}D_M$ to give $M = 12\text{-}B_M$.
From triangles $B_M D_M F_M$ and $A_M D_M E_M$

$$\frac{F_M D_M}{E_M D_M} = \frac{B_M F_M}{A_M E_M},$$

Whence,

$$\frac{x'}{c} = \frac{i'}{f},$$

$$x' = \frac{ci'}{f}.$$

Then,

$$M = l' - x' = l' - \frac{ci'}{f}.$$

In the given example:

$$M = 17.5 - \frac{5 \times 4}{9},$$

$$M = 17.5 - 2.2,$$

$$M = 15.3.$$

From the manner of derivation of the formulas, the two results are necessarily the same.

For the third quartile, Q_3, the required ordinate is erected at the base-line scale mark $3N/4$. With the given Table XLV and Fig. 43 this scale mark is $\frac{3 \times 24}{4} = 18$. By examination of the triangles $A_{Q_3}B_{Q_3}C_{Q_3}$ and $A_{Q_3}\,D_{Q_3}\,E_{Q_3}$, it is easily seen that

$$Q_3 = l + x = l + \frac{ci}{f},$$

Or from triangles $B_{Q_3}\,D_{Q_3}\,F_{Q_3}$ and $A_{Q_3}\,D_{Q_3}\,E_{Q_3}$

$$Q_3 = l' - x' = l' - \frac{ci'}{f}.$$

In the given table the eighteenth item falls in the fourth class. Then,

$$f = 4,\ l = 17.5,\ F_{Q_3} = 16,\ i = \frac{3N}{4} - F_{Q_3} = 18 - 16 = 2.$$

$$Q_3 = l + \frac{ci}{f} = 17.5 + \frac{5 \times 2}{4},$$

$$Q_3 = 17.5 + 2.5,$$

$$Q_3 = 20.0.$$

Or, from the upper end $N/4$ (= 6) items must be passed to reach Q_3. From column F' the sixth item from the largest is in the

third class from the bottom of the table. Then $f = 4$, $l' = 22.5$, $F'_{Q_3} = 4$, $i' = N/4 - F'_{Q_3} = 6 - 4 = 2$,

$$Q_3 = l' - \frac{ci'}{f} = 22.5 - \frac{5 \times 2}{4},$$

$$Q_3 = 22.5 - 2.5,$$

$$Q_3 = 20.0,$$

the same as from the other direction.

36. Formulas the Same for All Division Points.—The formulas derived for Q_3 are of exactly the same form as those for M. It is evident from the geometry of the figure that the same formulas hold for any division point whatsoever.

Let the student determine from both directions the value of Q_1 and of P_{80}.

37. Certain Division Points for Freshman Heights.—Suppose it be required to find the values of M, Q_1, Q_3, D_1, D_9, O_3, and P_{85} for freshman heights as given in Table XIV. The work may be arranged somewhat as in Table XLVI.

TABLE XLVI.—COMPUTATION OF DIVISION POINTS FOR FRESHMAN HEIGHTS

Division point	Horizontal scale mark	l	i	f	Computation
M	$\frac{N}{2} = 314\frac{1}{2}$	67.25	104.50	156	$67.25 + \frac{1.5 \times 104.5}{156} = 68.25$
Q_1	$\frac{N}{4} = 157\frac{1}{4}$	65.75	62.25	115	$65.75 + \frac{1.5 \times 62.25}{115} = 66.56$
Q_3	$\frac{3N}{4} = 471\frac{3}{4}$	68.75	105.75	148	$68.75 + \frac{1.5 \times 105.75}{148} = 69.82$
D_1	$\frac{N}{10} = 62.9$	64.25	18.9	51	$64.25 + \frac{1.5 \times 18.9}{51} = 64.81$
D_9	$\frac{9N}{10} = 566.1$	70.25	52.1	64	$70.25 + \frac{1.5 \times 52.1}{64} = 71.47$
O_3	$\frac{3N}{8} = 235\frac{7}{8}$	67.25	25.875	156	$67.25 + \frac{1.5 \times 25.875}{156} = 67.50$
P_{85}	$0.85N = 534.65$	70.25	20.65	64	$70.25 + \frac{1.5 \times 20.65}{64} = 70.73$

$N = 629$

38. Quartile Measure of Dispersion.—In the middle half of an array, that is, the portion between the first and third quartiles, the distribution is quite likely to be fairly uniform. This is especially true in an approximately normal distribution, since the classes of greatest frequency occur near the middle of the table. The problem of obtaining a measure of dispersion for approximately symmetric distributions then becomes very simple if only the middle half of the array is used. As the variation is

fairly uniform between the first and third quartiles, the size of the median will be very nearly equal to the arithmetic mean of the two quartiles.

$$M = \frac{Q_3 + Q_1}{2}.$$

The absolute values of the deviations of the two quartiles from the median will be nearly equal and will serve as a measure of dispersion. The *quartile measure of dispersion*, then, equals one-half of the interquartile range. It equals

$$M - Q_1 = Q_3 - M = \frac{Q_3 - Q_1}{2}.$$

39. Quartile Coefficient of Dispersion.—The corresponding coefficient of dispersion would be the measure of dispersion divided by that average from which the deviation was taken. *Quartile coefficient of dispersion* equals

$$\frac{Q_3 - Q_1}{2} \div M$$
$$= \frac{Q_3 - Q_1}{2} \div \frac{Q_3 + Q_1}{2}$$
$$= \frac{Q_3 - Q_1}{Q_3 + Q_1}.$$

The quartile measure of dispersion is very easily computed for distributions which can be put in array. In this respect it has an advantage over average deviation or standard deviation. In case of *normal* distribution, quartile deviation is about 0.67 of standard deviation. Quartile deviation is related to *probable error* to be considered later.

40. Quartile Deviation a Rough Measure.—Since one-fourth of the items are each less than the first quartile, one-fourth between the first quartile and the median, one-fourth between the median and the third quartile, and one-fourth larger than the third quartile, these three division points divide the entire set into four groups with the same number of items in each. The first quartile is the median of that half of the array smaller than the median of the whole array. The third quartile is the median of the other half of the array. Each is thus a type form for half of the array, and thus gives a rough idea of the distribution on each side of the median. So quartile deviation from the median gives a rough measure of dispersion of the entire group.

41. Frequency Classes Instead of Size Classes.—Instead of dividing the entire array of N items into four groups of $N/4$ each, a slightly more minute division may be made into eight groups of $N/8$ items each, or ten groups of $N/10$ items each. To obtain a still more minute analysis, the entire set is divided into 100 groups of $N/100$ items each. The division points are *octiles, deciles,* and *percentiles,* respectively. The result is that the N items are classified into groups of equal numbers (frequencies) each, instead of equal variations in size. For example, plotting the values of the deciles of freshman heights computed from

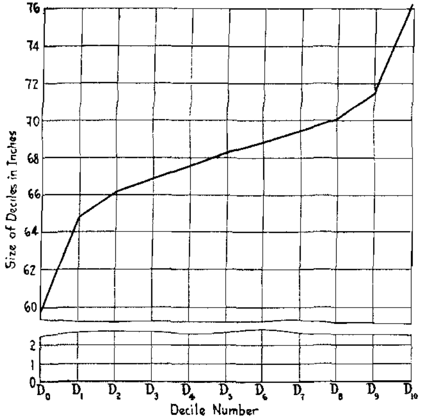


FIG. 44.—Deciles of freshman heights.

Table XIV, Fig. 44 is obtained. The plotted values are connected by straight lines.

DECILES OF FRESHMAN HEIGHTS

D_1	64.81
D_2	66.22
D_3	66.97
D_4	67.65
D_5	68.31
D_6	68.86
D_7	69.50
D_8	70.14
D_9	71.47

This graph shows that the variation from decile to decile is very uniform except at the two extremes of the table.

42. Division Points Graphically.—Any or all of the various division points may be obtained graphically from the ogive. For the pth percentile, lay off from the base line $1/p$th part of the longest ordinate to the curve. At this point draw a parallel to the base line. Where this parallel meets the curve, drop a perpendicular to the base line. At the point where it meets the base line, read the value of the pth percentile on the base line scale. Since any ordinate to the ogive measures the cumulative frequency up to the point where that ordinate stands, the ordinate as above constructed locates a division point such that p per cent of the items are less than the value at that point and all the rest are greater. The ogive is thus the most convenient means of getting a set of division points graphically.

43. Comparison of Measures of Dispersion.—Three distinct measures of dispersion have been studied. The question naturally arises as to which one to use in any given case. If the distribution is nearly symmetrical or only a rough estimate is needed, the quartile deviation does very well. If, however, the distribution is quite unsymmetrical, or if there are violent fluctuations in frequency from class to class, the quartile deviation is not very representative. Then average deviation would be better because it takes into account all deviations from the average used. Standard deviation also takes all deviations into account and is highly susceptible to algebraic manipulation. Standard deviation is derived from the arithmetic mean and would not be appropriate if some other average is used. The arithmetic mean is the most commonly used average and its associate, standard deviation, the most commonly used measure of dispersion. Especially is this true where mathematical processes are involved. Standard deviation gives more weight to extreme items than do the other measures of dispersion. Sometimes in statistics of business or of economics, it is important to bring out the influence of these extreme times. If the chief interest lies in that which is characteristic rather than in the exceptional case, the average deviation is likely to be more useful. If it is desired to shut out the exceptional cases entirely, then the quartile deviation may be used. In order to form a complete judgment of the situation, it is well to compute all three measures of dispersion. These, compared with the average and with each other, give a very comprehensive view.

In comparing dispersions of two or more variables having different averages, do not forget that the comparison should be made in terms of coefficient of dispersion instead of measure of dispersion.

44. Lorenz Curve.—One graphical means of showing dispersion is the Lorenz curve. It is usually applied to distribution of wealth, wages, income, and similar distributions, though it may be applied to any frequency distribution. The method is best shown by an illustrative example from Table XLVII.

TABLE XLVII.—DISTRIBUTION OF WEALTH IN CITY A

Classes, in thousands of dollars	Number of persons, in thousands	Amount of wealth, in millions of dollars
0– 4	120	300
5– 9	50	375
10– 24	22	385
25– 49	5	187.5
50– 99	2	150
100–149	1	125
Total.................	200	1,522.5

TABLE XLVIII.—DISTRIBUTION OF WEALTH IN CITY A
Cumulative

Classes, in thousands of dollars	Cumulative number of persons, in thousands	Cumulative amount of wealth, in millions of dollars
0– 4	120	300
5– 9	170	675
10– 24	192	1,060
25– 49	197	1,247.5
50– 99	199	1,397.5
100–149	200	1,522.5

From Table XLVII a cumulative table, Table XLVIII, of number of persons and amount of wealth is constructed. Table XLIX is made by reducing the cumulative columns of Table XLVIII to percentages.

TABLE XLIX.—CUMULATIVE PERCENTAGES OF NUMBER OF PEOPLE AND AMOUNT OF WEALTH POSSESSED BY THEM IN CITY A

Cumulative per cent, number of persons	Cumulative per cent, amount of wealth
60	20
85	44
96	70
98.5	82
99.5	92
100	100

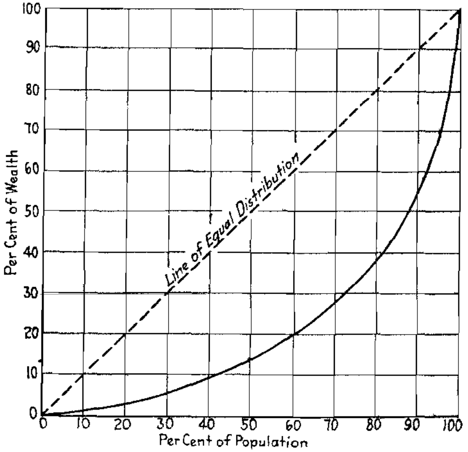


FIG. 45.—Lorenz curve. Distribution of wealth.

Table XLIX shows the percentage of wealth belonging to different percentages of the people and the percentage of population having various percentages of wealth. The data of Table XLIX are plotted and the points connected by straight lines or smoothed by a curve. Figure 45 shows a smooth curve through the points plotted. The curve starts with 0 per cent of population having 0

per cent of wealth. It is a long jump to the next point, 60 per cent of the population with 20 per cent of the wealth. In drawing this part of the curve, it would be necessary to judge its position by the form and direction of the right-hand part of the curve or else obtain data giving smaller class intervals for the percentages of the poorer portion of the people. The curve as drawn is probably nearer the truth than straight lines connecting the points would be.

45. Interpretation of Lorenz Curve.—Assuming this continuous curve to be true throughout its length, the per cent of wealth owned by any given per cent of the population, or the per cent of people owning any specified portion of the wealth, may be read from the graph. For instance, the poor half of the population has about 14 per cent of the wealth, as shown by the ordinate at 50 on the horizontal scale. The richest 5 per cent of the population has about 33 per cent of the wealth. This is determined by reading the ordinate at 95 per cent on the horizontal scale and subtracting the result from 100. That is, at 5 per cent less than 100 per cent of population, the curve is 33 per cent below 100 per cent of the wealth. The so-called middle class has what per cent of the wealth? This may be determined by computation from the figures read in the two cases above or may be taken directly from the graph. On the graph, take off with a divider the difference of ordinates at 50 and 95 per cent of the population. Transfer this difference to the per cent of wealth scale and read the answer, 53 per cent.

Half of the wealth is in possession of what per cent of the richer people? At 50 per ecnt of wealth on the curve, read the abscissa on per cent of population scale. It reads about 88.5 per cent, 11.5 per cent less than 100 per cent. Then half of the wealth is in the possession of 11.5 per cent of the people. And so with any section of the wealth or of population, the percentages may be read from this graph.

The dotted diagonal is the line of equal distribution showing what the graph would be if wealth were equally distributed throughout the entire population. The farther the actual graph departs from this line the more unequal is the distribution, the greater the dispersion.

By plotting, in the same figure, Lorenz curves for different localities, the dispersion may be compared. Or, by plotting for the same locality Lorenz curves for different dates, it may at

once be seen whether distribution is becoming more or less uniform. The Lorenz curve gives a very striking comparison and is perhaps more easily interpreted by the layman than is either the ordinary frequency curve or the ogive.

Exercises

1. Using the data of Table XIV, find the average deviation of freshman heights from the arithmetic mean, making a table the form of Table XXXIX.

2. Find the average deviation of freshman heights from the median, making a table of the form of Table XL.

3. Find the average deviation of freshman heights from the modal height.

4. Find the average deviation of freshman weights from the modal weight, using the data of Table XIII.

5. Find the average deviation from the median of monthly precipitation at Seattle.

6. Find the average deviation from the arithmetic mean of monthly precipitation at Seattle.

7. Find the average deviation from the mode of monthly precipitation at Seattle.

8. Find the average deviation from the mode of monthly mean temperatures at Seattle.

9. Find the average deviation from the median of monthly mean temperatures at Seattle.

10. Find the average deviation from the arithmetic mean of monthly mean temperatures at Seattle.

11. Find the corresponding coefficient of dispersion for each of the measures of dispersion that have been found in Exs. 1 to 10 inclusive.

12. Get data of monthly mean temperatures and monthly precipitation from your nearest local weather bureau office and find measures of dispersion and coefficients of dispersion as done in Exs. 1 to 11 inclusive. Compare the results with results from the tables for Seattle.

13. Find, by the direct method, the standard deviation of freshman weights from the data in Table XIII. Use the form of Table XLI.

14. Find, by the short-cut method, the standard deviation of freshman weights. Check with the result of Ex. 13. Use the form of Table XLII.

15. Find the value of δ/σ for freshman heights and for freshman weights. Which one is nearest to the theoretical value of four-fifths? Was this to be expected?

16. Make a table, showing for freshman weights the same set of division points as shown for heights in Table XLVI.

17. Find the quartile measure and quartile coefficient of dispersion for freshman heights and for freshman weights.

18. Find the standard deviation, average deviation from the arithmetic mean, and quartile measure of dispersion of the leaf lengths shown in Table XXIII. Obtain the corresponding coefficients of dispersion.

19. Find the standard deviation, average deviation from the arithmetic mean, and quartile measure of dispersion of the leaf breadths shown in Table XXIV. Obtain the corresponding coefficients of dispersion.

20. Determine the average deviation from the median leaf length found in Ex. 27, Chap. IV. Find the corresponding coefficient of dispersion.

21. Determine the average deviation from the median leaf breadth found in Ex. 28, Chap. IV. Find the corresponding coefficient of dispersion.

22. Using the table constructed in Ex. 19, Chap. II, find standard deviation of leaf lengths. Use both the direct method and the short-cut method. Find the corresponding coefficient of dispersion.

23. Using the table constructed in Ex. 20, Chap. II, find standard deviation of leaf breadths. Use both the direct method and the short-cut method. Find the corresponding coefficient of dispersion.

24. Using the table constructed in Ex. 19, Chap. II, find the quartile measure and the quartile coefficient of dispersion of leaf lengths.

25. Using the table constructed in Ex. 20, Chap. II, find the quartile measure and the quartile coefficient of dispersion of leaf breadths.

26. Make a single table showing all the measures and coefficients of dispersion found in Exs. 20 to 25. Construct the table so that comparisons can be readily made. Which measure of dispersion of leaf lengths is greatest? Which least? Which coefficient of dispersion of leaf lengths is greatest? Which least? Answer the same questions in regard to leaf breadths. Which has the greater scatter, leaf lengths or leaf breadths? Which has the greater relative amount of scatter?

27.

TABLE A.—DISTRIBUTION OF LABOR INCOMES OF FARMERS, 1910 TO 1915[1]

Classes of annual labor income	Number in each income class	Total labor income, in millions
$-1,500 and less	64,000	$- 96
-1,500-$-1,000	76,800	- 96
-1,000- - 500	300,800	-226
- 500- 0	1,529,600	-382
0- 500	2,336,000	+584
500- 1,000	1,132,800	850
1,000- 1,500	473,600	592
1,500- 2,000	217,600	381
2,000- 2,500	96,000	216
2,500- 3,000	64,000	176
3,000- 4,000	64,000	224
4,000- 5,000	19,200	86
5,000- 10,000	25,600	192
Over 10,000	0	0
Totals	6,400,000	$2,501

[1] From "Income in the United States," vol. II.

TABLE B.—DISTRIBUTION OF LABOR INCOMES OF FARMERS, 1916 TO 1918[1]

Classes of annual labor income	Number in each income class	Total labor income in millions
$ −1,500 and less	84,500	$ −127
−1,500–$ −1,000	65,000	− 81
−1,000– − 500	266,500	−200
− 500– 0	903,500	−226
0– − 500	1,859,000	+465
500– 1,000	1,293,500	970
1,000– 1,500	708,500	886
1,500– 2,000	396,500	694
2,000– 2,500	299,000	673
2,500– 3,000	169,000	465
3,000– 4,000	201,500	705
4,000– 5,000	84,500	380
5,000– 10,000	117,000	877
Over 10,000	52,000	520
Totals	6,500,000	$6,001

[1] From "Income in the United States," vol. II.

Make additional columns to Tables A and B, providing the data for plotting a Lorenz curve for each table. Assume that the average income in each class is the midpoint of the class except for the first and last classes. For the first class of −$1,500 and less assume that the average for the class is −$1,500. For the last class of over $10,000 assume that the average for the class is $10,000.

Having completed the tables plot a Lorenz curve for each table, both on the same set of axes. Plot the 1910 to 1915 curve in red and the 1916 to 1918 curve in black.

In which period was the distribution of income of farmers more uniform? Does this show a greater degree of prosperity for farmers in that period?

What is the significance of the point where the curve crosses the line of 0 per cent of income?

If the tables were lost, how could you tell from the curve what per cent of the farmers received negative income?

28. AUSTRALIAN WAR CENSUS OF INCOMES[1]

Income class	Males		Females	
	Number	Amount of income to nearest thousand	Number	Amount of income to nearest thousand
Under £50	145,513	£ 4,163	301,592	£ 6,717
£ 50 and under £ 100	327,835	24,308	168,106	11,416
100 and under 150	448,195	55,090	52,929	6,250
150 and under 156	46,630	7,093	3,651	558
156 and under 200	157,350	27,219	12,697	2,211
200 and under 300	106,324	25,191	11,001	2,641
300 and under 500	49,108	18,388	6,617	2,498
500 and under 750	15,928	9,603	2,691	1,633
750 and under 1,000	6,313	5,393	1,145	970
1,000 and under 1,500	4,933	5,994	905	1,089
1,500 and under 2,000	2,132	3,676	364	629
2,000 and under 3,000	1,707	4,149	317	772
3,000 and under 4,000	659	2,249	102	361
4,000 and under 5,000	375	1,685	58	258
5,000 and over	746	7,300	86	656
Total................				

[1] From "Income in the United States," vol. II.

Make a table showing cumulative values and cumulative per cents both for number and for amount of income for males and for females. From the cumulative per cent columns plot points for a Lorenz curve for males. On the same set of axes plot points for a Lorenz curve for females. Smooth both graphs, using black for males and red ink for females. Place a proper legend in the lower right-hand corner.

Is there a marked difference shown in distribution of wealth for males and for females?

Is there more or less unequal distribution for females than for males? What shows this?

What per cent of total wealth of each sex is possessed by the poorer half of that sex?

What per cent of the total wealth of each sex is possessed by the richest 5 per cent of that sex?

What per cent of the total wealth of each sex is possessed by the so-called "middle half," *i.e.*, by the portion between 50 and 95 per cent?

CHAPTER VI

SKEWNESS

It was noted in the case of freshman heights that the frequency distribution was nearly symmetrical. This was evident from the smoothed frequency curve of Fig. 31.

1. Skewness.—In the case of freshman weights there was a lack of symmetry. The smooth curve of Fig. 29 tails off to the right of the ordinate to the high point farther than it does to the left. When it happens that the curve tails off farther to one side of the mode than to the other, *skewness* is said to be present. The distribution and the curve are said to be *skewed.*

When the values of the variable are a set of observations on a single thing and there is no reason for thinking that positive deviations from the arithmetic mean are any more or any less likely to occur than negative deviations of the same size, a symmetrical distribution may be expected. Such an example would be a large number of measurements of the same line by a surveyor. But if the observations consist of a set of measurements of a single characteristic of a large number of objects, a certain degree of skewness is likely to be found. The measurements of the weights of 629 freshmen constituted such a set. So also did the measurements of the heights of these freshmen. In one case skewness was noticeable; in the other case, not.

It is an important fact that in many cases skewness is sufficiently slight to be neglected.

2. Effect on Mode, Median, and Arithmetic Mean.—If the frequency curve tails off farther to the right of the mode than to the left, it will almost invariably follow that more than half the area under the curve falls to the right of the ordinate at the mode. Since the ordinate at the median bisects the area under the curve, the median must then be shifted to the right and be larger than the mode. The ordinate at the arithmetic mean passes through the center of gravity of the area under the curve. This will also almost invariably be to the right of the maximum ordinate if the curve tails off farther to the right than to the left. Then the arithmetic mean will be larger than the mode. The

median is shifted to the right only by virtue of an increased number of large items over what there would be if the part of the curve to the right of the mode were symmetrical with the part to the left. The center of area is shifted only because of added area to the right. In getting the position of the center of gravity each element of new area is multiplied by its distance from the vertical axis and added to the previous products. The greater the distance from the axis, the greater the added product, and hence the greater effect in shifting the center of gravity. It follows that almost invariably the ordinate through the center of gravity of the area will be shifted farther from the maximum ordinate than is the ordinate through the center of area. In other words, in a skewed distribution tailing off farther to the right than to the left, the difference between the arithmetic mean and the mode is almost invariably greater than the difference between the median and the mode. If the distribution tails off farther to the left of the ordinate at the mode than to the right, the same reasoning leads us to expect the same relationship of differences among mode, median, and arithmetic mean. It has been found empirically that in many cases

$$(a - Z) = 3(a - M).$$

In the case of freshman weights

$$a - Z = 4.5,$$
$$a - M = 1.4,$$

and the above relation holds nearly. In many cases, however, the relation does not hold.

3. Measure of Skewness.—A natural measure of skewness would be a number determined by the effect of the skewness. When there is symmetry, there is no skewness. Then arithmetic mean, median, and mode are all equal. Skewness separates these three statistical constants. The greater the skewness, the greater the separation, usually. The amount the arithmetic mean is separated from the mode, that is, the deviation of the arithmetic mean from the mode,

$$a - Z,$$

gives a measure of skewness. The deviation of the median from the mode,

$$M - Z,$$

may also be used. So may

$$a - M,$$

the deviation of the arithmetic mean from the median.

4. Coefficient of Skewness.—Suppose that $a - Z$ for one variable equals $a - Z$ for another variable, but that the variability, dispersion, scatteration, in one case is greater than in the other. It is evident that the deviation of arithmetic mean from the mode is relatively less in the case of large dispersion than in the case of small dispersion. So for a coefficient of dispersion $a - Z$ should be divided by the measure of dispersion, either standard deviation σ, or average deviation δ_a. Coefficient of dispersion is

$$\frac{a - Z}{\sigma},$$

or

$$\frac{a - Z}{\delta_a}$$

The deviation of arithmetic mean from the mode seems to be the most logical of the above measures, but it frequently happens that the mode is not well defined. In that case $a - M$ makes a better measure for skewness. The corresponding coefficient is either

$$\frac{a - M}{\sigma}$$

or

$$\frac{a - M}{\delta_a}.$$

5. Algebraic Sign.—If $a > Z$, skewness is positive, and the frequency curve usually tails off farther to the right than to the left. If $a < Z$, skewness is negative and the frequency curve usually tails off farther to the left.

6. Moments.—Another measure of skewness sometimes used is the cube root of what is known as third moment about the arithmetic mean. Moments about the arithmetic mean are defined as follows:

$$\textit{First moment} = \frac{\Sigma d_a}{N},$$

$$\textit{Second moment} = \frac{\Sigma d_a^2}{N},$$

$$\textit{Third moment} = \frac{\Sigma d_a^3}{N},$$

and so on for higher moments.

It might seem in the case of positive skewness with large positive values of d_a and small negative values that first moment might measure skewness. It does not, because

$$\frac{\Sigma d_a}{N} = 0,$$

since Σd_a always equals 0.

In second moment about the arithmetic mean, squares of the deviations are used and an excess of positive or of negative deviations would not be revealed. So

$$\frac{\Sigma d_a^2}{N}$$

is not a measure of skewness. Square root of second moment about the arithmetic mean is standard deviation.

7. Third Moment a Measure of Skewness.—With third moment about the arithmetic mean, however, an excess of positive deviations will make Σd_a^3 positive and an excess of negative deviations will make it negative. Deviations being cubed, large deviations will have great weight as compared with small ones. So that the greater the skewness, the greater will be the third moment, positive if skewness is positive, and negative if skewness is negative. As square root of second moment about the arithmetic mean was used to give standard deviation, so here cube root of third moment about the arithmetic mean is used for the measure of skewness.

$$Skewness = \sqrt[3]{\frac{\Sigma d_a^3}{N}}.$$

The corresponding coefficient of skewness would naturally be

$$\frac{\sqrt[3]{\dfrac{\Sigma d_a^3}{N}}}{\sqrt{\dfrac{\Sigma d_a^2}{N}}} = \frac{\text{cube root of third moment}}{\sigma}.$$

This measure and coefficient are laborious to compute. However, in the problem of fitting a mathematical curve to a given frequency distribution, as high as the fourth moment is used. Professor Karl Pearson has devised a measure of skewness involving moments as high as the fourth.

8. Quartile Measure of Skewness.—As was done for measure of dispersion, so with skewness, only the middle half of the array

may be used if one wishes to eliminate the influence of the extreme items. If there is symmetry of distribution

$$M - Q_1 = Q_3 - M.$$

Suppose the symmetry is destroyed by adding n large items to the array. Addition of items at the right-hand end of the array is sure to shift the positions of both quartiles and the median to the right. Q_1 is shifted $n/4$ items; M, $2n/4$ items; and Q_3, $3n/4$ items to the right. Consideration of the symmetrical frequency curve shows that Q_1 is shifted to a region of greater frequency. A movement over $n/4$ items will then change the size of Q_1 but little. M is shifted to a region of less frequency though the frequency is not diminishing rapidly. A movement over $2n/4$ items will change the size of M more than the size of Q_1 was changed. Q_3 is shifted to a region of less frequency in a part of the distribution where frequeneies are diminishing more rapidly. A movement over $3n/4$ items will change the size of Q_3 more than the size of M was changed. In consequence of these changes

$$M - Q_1 < Q_3 - M.$$

The difference between Q_3 and M has become greater than the difference between Q_1 and M in consequence of the skewness that has been introduced. The difference of these differences then gives a measure of skewness in which the influence of the size of extreme items has been eliminated. The *quartile measure of skewness* then equals

$$(Q_3 - M) - (M - Q_1) = Q_3 + Q_1 - 2M.$$

If skewness has become negative instead of positive, the same reasoning would have applied. $M - Q_1$ would be greater than $Q_3 - M$, and the resulting measure of skewness would be negative. A quartile coefficient of skewness may be obtained by dividing by the entire interquartile range. *Quartile coefficient of skewness* equals

$$\frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1}.$$

All these measures and coefficients of skewness become zero, as they should, when the distribution is symmetrical.

The probability distributions derived from the expansion of

$$(p + q)^n$$

are symmetrical for all values of n if, and only if, $p = q$. If $p \neq q$ the distribution becomes asymmetrical, approaching symmetry as n increases.

9. Skewness from Pure Chance.—The throwing of dice gives a good example. If one die is thrown, the probability that an ace will come up is $\frac{1}{6}$, and the probability that an ace will not come up is $\frac{5}{6}$. The same theorems on probability that were applied to the result of tossing coins apply here to show that the terms of the expansion of

$$(\tfrac{5}{6} + \tfrac{1}{6})^n$$

give, respectively, the probabilities of throwing no ace, one ace, two aces, and so on, in throwing n dice once. If the n dice are thrown N times, then N times each of these terms gives the corresponding expected frequencies of no ace, one ace, two aces, and so on.

10. Example.—Suppose seven dice are thrown 1,000,000 times.

$$\left(\frac{5}{6} + \frac{1}{6}\right)^7 = \frac{1}{6^7}(5^7 + 7 \cdot 5^6 \cdot 1 + 21 \cdot 5^5 \cdot 1^2 + 35 \cdot 5^4 \cdot 1^3 + 35 \cdot 5^3 \cdot 1^4 + 21 \cdot 5^2 \cdot 1^5 + 7 \cdot 5 \cdot 1^6 + 1^7)$$

$$= \frac{1}{279{,}936}(78{,}125 + 109{,}375 + 65{,}625 + 21{,}875 + 4{,}375 + 525 + 35 + 1).$$

Multiplying each term by 1,000,000 gives the expected frequencies as shown in Table L.

TABLE L.—PROBABLE FREQUENCIES OF ACES IN 1,000,000 THROWS OF SEVEN DICE

Number of aces	Frequency
0	279,000
1	391,000
2	234,000
3	78,300
4	15,700
5	1,870
6	126
7	4
Total	1,000,000

Inspection of Table L, without plotting a curve, shows a fair degree of symmetry if only the first three entries are considered. Beyond that point, the distribution tails off rapidly to very few of the largest number of aces possible.

11. Measures of Skewness of Freshman Weights.—Referring to the measurements of freshman weights,

$$a = 142.25.$$
$$Z = 137.8.$$
$$M = 140.5.$$
$$Q_1 = 129.4.$$
$$Q_3 = 153.2.$$
$$\sigma = 17.9.$$

Measure of skewness,

$$a - Z = 4.4.$$

Coefficient of skewness,

$$\frac{a - Z}{\sigma} = 0.24.$$

Quartile measure of skewness,

$$Q_3 + Q_1 - 2M = 1.6.$$

Quartile coefficient of skewness,

$$\frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1} = 0.07.$$

12. Significant Degree of Skewness.—The use of certain statistical processes depends on the distribution being normal, or approximately so. Skewness measures the degree of lack of symmetry. The question as to how great skewness may be before these processes cease to have significance is not fully settled. It is safe to say, however, that if the measure of skewness is more than four times its probable error, the distribution is sufficiently far from normal distribution to destroy the reliability of these processes. The meaning of "probable error" will be discussed presently.

There are many distributions, particularly among biological measurements, which follow the normal distribution approximately. In economics the normal distribution is much less likely to occur. It also frequently happens that a distribution is very much skewed, but the logarithms of the values of the variable have nearly a normal distribution. This gives a point of attack for studying the distribution.

13. Extreme Forms.—There are extreme types of asymmetry. The greatest frequencies may occur with the very smallest sizes and decrease as the size of item increases. Or, the small sizes of the variable may have very low frequencies increasing with increase of the size of the variable. For these types the frequency

curve does not go up to a high point and then descend, but is high at one end and low at the other. Such distributions are said to be "*J-shaped.*" There are cases in which high frequencies occur both for small sizes and for large sizes of the variable, with low frequencies for the medium sizes. The curve is high at both ends and low in the middle. Such distributions are called "*U-shaped.*" Examples of J- and U-shaped distributions are given by Yule.[1]

Exercises

1. Find $a - Z$ for freshman heights. Using this as a measure of skewness, which has the greater skewness, the distribution of freshman heights, or of freshman weights? In making the comparison, should the measure of skewness be used or the coefficient of skewness? Using $a - Z$ as a measure of skewness, find the coefficient of skewness of freshman heights. Compare this with the coefficient of skewness of freshman weights.

2. Find the quartile measure of skewness of freshman heights. How does this compare with the corresponding measure for freshman weights? Would it be expected that these measures would be larger or smaller than the corresponding $a - Z$ measures? Why?

3. How do the quartile coefficients of skewness of freshman heights and freshman weights compare?

4. Find $a - Z$ for the leaf lengths of Ex. 19, Chap. II, and $a - Z$ for the leaf breadths of Ex. 20, Chap. II. Using $a - Z$ as a measure of skewness, which has the greater skewness, the distribution of leaf lengths or that of leaf breadths?

5. Find the quartile measure of skewness for leaf lengths of Ex. 19, Chap. II, and for leaf breadths of Ex. 20, Chap. II.

[1] G. Udney Yule, "An Introduction to the Theory of Statistics," pp. 98–105.

CHAPTER VII

PROBABLE ERROR

The meaning of the term "probable error" will now be ascertained.

1. Definition of Probable Error.—The interquartile range $(Q_3 - Q_1)$ includes exactly half the items in the array. Suppose each value of the variable is written on a card and placed in a sack. Shake up the sack and draw a card at random. The chances are even that the value on the card drawn will be inside the interquartile range, since exactly half the values are within this range and half outside. The probability of drawing a value within the interquartile range is $\frac{1}{2}$. If, now, the distribution follows the normal probability curve,

$$Q_3 - M = M - Q_1$$

equals the semi-interquartile range. Moreover,

$$M = a,$$

and the semi-interquartile range is the quartile deviation from the arithmetic mean. This quartile deviation from the arithmetic mean is known as the probable error of the series of values of the variable. Probable error of a set of observations may then be defined as *that deviation from the arithmetic mean on either side within which exactly half the observations lie.* If one of the observations be selected at random, it is as likely as not that it will differ from the arithmetic mean by not more than the probable error.

It can be shown that

$$p.e. = 0.67\sigma$$

nearly.

2. x/σ and Normal Probability Curve.—In the discussion of symmetric distributions in Chap. III, pages 77 to 83, x/σ was used as the unit of measure on the axis of the independent variable. Let it be recalled that x is deviation from the arithmetic mean and σ is standard deviation. The equation of the normal probability curve has been derived and found to be

$$y = y_0 e^{-\frac{x^2}{2\sigma^2}}$$

as was referred to on page 82, Chap. III. The areas under the curve from the y-axis up to any ordinate have been tabulated for varying values of x/σ.[1] The y-axis goes through the highest point of the curve and so is at the mode. From the symmetry of the curve, mode equals arithmetic mean equals median. So the y-axis is at the arithmetic mean and at the median. Then x, being measured from the arithmetic mean is deviation from arithmetic mean. The area under the curve from the left-hand end up to any ordinate is the total frequency up to the size of item at that ordinate. The area up to the y-axis is half of the total area. The area from the y-axis up to the third quartile must be one-quarter of the entire area, since one-quarter of the total frequency is included between the median and the third quartile.

3. Areas under Normal Probability Curve.—Table LI is made from a table of areas under the normal probability curve. The entire area is taken as 1. If $N = 100{,}000$, then 100,000 times the entries in the second column gives the total frequency up to the ordinate at x/σ. Similarly for the other columns.

TABLE LI.—AREAS UNDER THE NORMAL PROBABILITY CURVE

$\frac{x}{\sigma}$	Fraction of area up to the ordinate at $\frac{x}{\sigma}$	Fraction of area from y-axis up to ordinate at $\frac{x}{\sigma}$	Fraction of area between the ordinates at $\frac{x}{\sigma}$ and $-\frac{x}{\sigma}$ α	$1 - \alpha$
0	0.50000	0	0	1
0.67449	0.75000	0.25000	0.50000	0.50000
1	0.84134	0.34134	0.68269	0.31731
2	0.97725	0.47725	0.95450	0.04550
3	0.99865	0.49865	0.99730	0.00270
4	0.99997	0.49997	0.99994	0.00006

x/σ = unit of measure on the x-axis.

One-fourth of the entire area, 0.25000, from the y-axis is at

$$\frac{x}{\sigma} = 0.67449$$

[1] JAMES W. GLOVER, "Tables of Applied Mathematics." KARL PEARSON, "Tables for Statisticians and Biometricians." Table II gives the entire area from the left end up to the ordinate in question. The tabular values may be obtained by a process known as integration between limits.

and

$$x = 0.67449\sigma.$$

As this value of x is a quarter of the entire frequency from the middle of the array, it is the value of the deviation of the third quartile from the arithmetic mean which here is also its deviation from the median.

By definition, the deviation of third quartile from the arithmetic mean is probable error. Hence the relation,

$$p.e. = 0.67\sigma,$$

nearly. This is semi-interquartile range.

4. Skewed Distribution.—If the distribution is skewed, the deviation of the third quartile from the arithmetic mean is no longer the semi-interquartile range. In such case probable error is simply defined as being

$$p.e. = 0.67449\sigma.$$

Thus $p.e.$ simply becomes a measure of variation equal to about two-thirds of standard deviation.

Always exactly half the observations lie between the first and third quartiles. Only in case of normal probability distribution do exactly half the observations lie between $x = -p.e.$ and $x = +p.e.$ Values of a large number of measurements of a single object are likely to follow normal probability distribution quite closely. No significant error would then be expected in assuming that half the observations lie between arithmetic mean minus probable error and the arithmetic mean plus probable error.

5. Chances of a Random Selection Falling beyond Certain Limits.—In the last column of Table LI is tabulated the fraction of the total frequency that will fall outside the limits of $-x/\sigma$ to $+x/\sigma$. In the next to the last column is tabulated the fraction of the total frequency that falls between the limits $-x/\sigma$ to $+x/\sigma$. At $x/\sigma = 0.67449$ half the items fall outside these limits and half within. So in choosing an item at random the chances are 1:1 of its falling between these limits. At $x/\sigma = 1$ there are 0.31731 of the items outside of $x = -\sigma$ to $x = +\sigma$, and 0.68269 of the items between the limits $x = -\sigma$ to $x = +\sigma$. Since 0.31731 items are about one-third of the total, in choosing an item at random, the chances of getting one differing from the arithmetic mean by not more than σ are about 2:1. In the same way, at $x/\sigma = 2$, 0.04550 of the items fall outside the interval from $x = -2\sigma$ to $x = +2\sigma$. This is about one-twenty-fifth of the total number of items. So

in choosing an item at random from the entire set, the chances are about 24:1 *of getting one differing from the arithmetic mean by* more than 2σ. Following on down the column, it is seen that the chances of getting an item at random with deviation from arithmetic mean of more than 4σ are less than 16,000:1. In other words only about 1/16,000 of the entire area under the curve is not included between the ordinates at $x = -4\sigma$ and $x = +4\sigma$. *A deviation of more than 4σ can scarcely happen by influence of chance alone.*

The size of standard deviation may thus be used as a means of measuring the degree of reliability of a set of observed values of a variable. Custom has introduced the convention of using probable error instead of standard deviation. Since

$$p.e. = 0.67449\sigma,$$

it is as definite a measure as is standard deviation itself. Probable error has the simplicity of being that deviation from the arithmetic mean such that, if an item is selected at random from the entire set, the chances are even that its deviation from the arithmetic mean will not be more than the probable error. It must be remembered that this statement is strictly true only in case of a normal probability distribution.

Now tabulate areas under the normal probability curve in terms of probable error instead of standard deviation, as shown in Table LII.

TABLE LII.—AREAS UNDER THE NORMAL PROBABILITY CURVE

$\frac{x}{p.e.}$ = unit of measure on the x-axis

$\frac{x}{p.e.}$	Fraction of area from y-axis to $\frac{x}{p.e.}$	Fraction of area between the ordinates at $-\frac{x}{p.e.}$ and $+\frac{x}{p.e.}$ A	Fraction of area outside the ordinates at $-\frac{x}{p.e.}$ and $+\frac{x}{p.e.}$ $1 - A$	Odds against the occurrence of a deviation greater than x $A \div (1 - A)$
1	0.25000	0.50000	0.50000	1:1
2	0.41131	0.82262	0.17738	4.6:1
3	0.47848	0.95696	0.04304	22:1
4	0.49651	0.99302	0.00698	142:1
5	0.49962	0.99924	0.00076	1,300:1
6	0.49997	0.99995	0.00005	20,000:1

6. The Limits in Terms of Probable Error.—Examine the meaning of these figures at, say, $x = 3$ $(p.e)$. Let $N = 100{,}000$. The total frequency from the arithmetic mean up to a deviation from the arithmetic mean of three times the probable error is 47,848. The total frequency included between deviations of three times the probable error on each side of the arithmetic mean is 95,696. The total frequency outside these limits is 4,304. If, then, an item is selected at random from the entire set, the probability of getting one differing from the arithmetic mean by more than three times the probable error is $4{,}304 \div 100{,}000 = 0.04304$, or about 4.30 in 100. The odds against getting an item having a deviation from the arithmetic mean of more than three times the probable error are $95{,}696 \div 4{,}304$, or about 22:1.

7. Probable Error of Various Constants.—Probable errors of various statistical constants have been computed by mathematical processes not necessary to reproduce here. Table LIII gives a number of such values.

TABLE LIII.—PROBABLE ERRORS OF VARIOUS STATISTICAL CONSTANTS

No.	Constant	Probable error
1	Arithmetic mean a of N observations	$0.6745 \frac{\sigma}{\sqrt{N}}$
2	Median M of N observations	$0.8454 \frac{\sigma}{\sqrt{N}}$
3	Standard deviation σ of a normal distribution	$0.6745 \frac{\sigma}{\sqrt{2N}}$
4	Coefficient of correlation r, normal distribution	$0.6745 \frac{1 - r^2}{\sqrt{N}}$
5	Correlation ratio η, approximately	$0.6745 \frac{1 - \eta^2}{\sqrt{N}}$
6	Any quantity in terms of standard deviation of that quantity	0.6745σ
7	Semi-interquartile range	$0.5306 \frac{\sigma}{\sqrt{N}}$
8	Difference between two averages $a_1 - a_2$, whose probable errors E_1 and E_2 are known.	$\sqrt{E_1^2 + E_2^2}$

In explanation of these formulas take, say, No. 6. If N observations are made on the value of a quantity, the standard deviation may be computed. The probable error is 0.6745 times the computed standard deviation. Take No. 1 as another illustration. If the arithmetic mean of each of a number of sets of N

observations each are determined, this set of means has a standard deviation and probable error. This is the probable error of the arithmetic mean of a set of N observations and is found to be $0.6745\sigma/\sqrt{N}$.

8. Examples.—The probable error of a quantity is written immediately after the value of the quantity with a $\pm$ sign between them. The probable error of the arithmetic mean of freshman heights is

$$0.6745 \frac{2.51}{\sqrt{629}} = 0.067.$$

The arithmetic mean of freshman heights would then be recorded as

$$a = 68.18 \pm 0.067.$$

This means that if a large number of groups of 629 freshmen each were measured and the arithmetic mean height of each group were computed, one would expect to find about half of these means to fall between 68.18 − 0.07 = 68.11 and 68.18 + 0.07 = 68.25. The chances are even that the correct arithmetic mean differs from 68.18 by not more than 0.07.

The chances are very small that the true arithmetic mean differs from 68.18 by more than 0.27, or four times the probable error. In other words, if many similar samples of freshman heights were taken, the chances are very small of any arithmetic mean height being less than 68.18 − 0.27 or greater than 68.18 + 0.27, that is, less than 67.91 or more than 68.45 in.

A good illustration of the practical application of probable error is given by Professor Pearl.[1] Suppose the average pulse rate of 150 people is found to be 79.68 ± 0.15 beats per minute. After administering a certain drug to each person, the average pulse rate is found to be 81.12 ± 0.20 beats per minute. May this increase of pulse rate be regarded as quite surely due to the drug or may it be a result of chance due to the sample?

The difference of the two averages is

$$a_1 - a_2 = 81.12 - 79.68 = 1.44.$$

The probable errors are $E_1 = 0.15$ and $E_2 = 0.20$. From No. 8 in Table LIII

$$p.e.\ (a_1 - a_2) = \sqrt{(0.15)^2 + (0.20)^2} = 0.25.$$

[1] RAYMOND PEARL, "Medical Biometry and Statistics," p. 214.

The difference of the averages is then written

$$a_1 - a_2 = 1.44 \pm 0.25.$$

While the difference of 1.44 beats per minute is not very great, yet it is nearly six times its probable error. From Table LII the odds against a deviation six times the probable error occurring from chance alone are about 20,000: 1. In other words, the chance of a difference of

$$1.44 \pm 0.25$$

heart beats per minute not being significant is only about 1 in 20,000. This makes it very certain that the drug administered had significant effect in increasing the average number of heart beats per minute.

The meaning of the statement on page 175 in regard to skewness being more than four times its probable error should now be understood. Unless a deviation is more than four times its probable error, there is no great reason for thinking it may not result from pure chance.

9. Value of Probable Error Depends on N.—From each of the formulas given in Table LIII, it is seen that the greater N is the smaller the probable error. So the reliability of a result, other things being equal, increases with increase of the number of observations made, when the distribution is approximately normal. The dividing line between large and small numbers of observations is indefinite. The problem of distribution in small samples is an important one.[1]

10. Use of Probable Error.—Probable error serves as a measure of the extent to which a quantity may deviate from the average, due to chance alone. It thus becomes of great importance in connection with all observed measurements.

Exercises

1. Find the probable error of the aritbmetic mean freshman height and the standard deviation of freshman height. Are these probable errors sufficiently large to make the arithmetic mean or standard deviation unreliable?

2. If another group of 629 freshmen were selected, how much might the arithmetic mean freshman height of this group differ from the arithmetic mean of the sample of 629 used in this book without making it reasonably certain that this new group should not be used for studying average height of the freshman population of which our sample is representative?

[1] See Karl Pearson, "On the Distribution of Standard Deviations of Small Samples," *Biometrika*, vol. 10, 1915.

3. Find the probable error of the arithmetic mean of mean monthly temperatures at Seattle, and the probable error of the standard deviation of monthly precipitation at Seattle.

4. Find the probable error of the arithmetic mean of the number of heads per throw of the seven dimes of Table XV.

5. Find the probable error of the arithmetic mean leaf length determined from the table of Ex. 19, Chap. II.

6. Find the probable error of the arithmetic mean leaf breadth determined from the table of Ex. 20, Chap. II.

CHAPTER VIII

CURVE FITTING

In studying frequency graphs an attempt was made to draw a smooth curve which was judged to be fairly representative of the frequency distribution of the entire universe from which the sample had been selected. In other words, the attempt was made graphically to pass from the properties of the random sample to the corresponding properties of the universe from which it was selected. It was discovered in the case of heights of freshmen that the frequency polygon fell fairly close to a symmetric curve determined by the terms of the expansion of

$$N(\tfrac{1}{2} + \tfrac{1}{2})^{11}.$$

Thus was found a definite mathematical function which followed quite closely the variations in freshman heights. Moreover, the polygon determined by the terms of the expansion of

$$N(\tfrac{1}{2} + \tfrac{1}{2})^{n}$$

approaches a definite curve as a limit as n increases without limit. This curve, the normal probability curve, gives an equational relationship between the independent variable x and the function y. This equation,

$$y = y_0 e^{-\frac{x^2}{2\sigma^2}}$$

enables us to determine y for any given value of x. It will be recalled that x stands for deviation from the arithmetic mean, and y is the corresponding frequency in a normal probability frequency distribution. From the above discussion it would seem that this equation could be used to give a fair approximation to the frequency to be expected for any given freshman height among a universe of freshman heights for which the given sample is representative.

In the case of freshman weights, the frequency curve as drawn did not prove so nearly symmetrical. There should be some means of deciding whether a normal probability curve fits

sufficiently close for practical purposes. Perhaps a curve determined by the terms of the expansion of

$$N(p + q)^n,$$

where p does not equal q, could be found which would fit more closely. If it is possible to get a mathematical relationship between two variables, then a better analysis may be made than if there is no apparent relationship.

1. Curve of Closest Fit.—This leads at once to the idea of determining the curve of closest fit to a given set of data, of such a nature that it may be expected fairly to represent the entire universe from which the data are selected. The kind of curve to use in any given instance must be left largely to that good judgment which is the result of experience. When a set of points is plotted from a given set of observed data, it is quite likely to be noticeable that these points tend to arrange themselves in a definite manner. In other words, there appears to be some definite law governing the distribution of the points. Certain considerations help in selecting the type of curve whose equation is a statement of this law. It is then necessary to determine that particular curve of this type which is the curve of closest fit to the points plotted. Probably the curve will not pass through all the points plotted, due to the accidental errors of observation or the accidental deviations in selecting a random sample. The idea of the curve is to bring out the general law, independent of accidental variations. It is always possible to write the equation of a curve passing through all the points. Its equation will, in general, have as many constants as there are points through which the curve is to pass. This not only makes the equation unwieldy, if there are many points, but destroys the very purpose of bringing out the general tendencies as shown by the required law.

2. Certain Type Forms.—Experience has shown that in most cases the laws of nature can be expressed by means of a few type forms of functions with relatively few constants involved. The straight line and the parabola are two of the forms frequently occurring. The equation of the straight line is

$$y = ax + b.$$

That of the parabola is

So-called parabolas of the nth degree with equations of the form

$$y = ax^n,$$

and the exponential function with equation of the form

$$y = ae^{bx},$$

are also not infrequent. These four types will now be considered. There may also be mentioned other second-degree equations which include all conic sections, different forms of trigonometric equations, logarithmic equations, and equations which are combinations of the functions given.

3. Straight Line.—The straight line is one of the simplest forms to fit to a given set of data. There are several ways of getting the straight line of closest fit. If only two points are given, a straight line can always be passed through them by substituting the coordinates of each point for x and y in the equation

$$y = ax + b$$

and solving the two resulting equations for a and b. The values of a and b thus found, substituted in the original equation, give the equation sought. For example, suppose that for $x = 1, y = 3$, and for $x = 5, y = 11$.

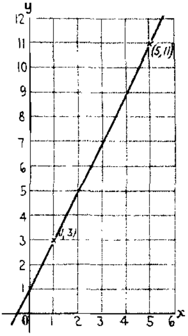


Fig. 46.—Straight line through two points.

Then the two equations are

$$3 = a + b$$
$$11 = 5a + b.$$

Solving for a and b gives

$$a = 2, b = 1.$$

The required equation is, then,

$$y = 2x + 1.$$

The graph is shown in Fig. 46.

4. Graphical Method.—*Take the following values of x and y:*

x	0	5	10	15	20	25	30	35	40
y	10	15	20	27	28	35	41	44	50

Plot the points as shown in Fig. 47. A thread may now be stretched along the plotted points and shifted until it seems to be in the position of closest fit to the points. When so placed, its position is marked by a straight line. This will be the line

required. This may be called the *graphical method*. It requires good judgment in placing the thread.

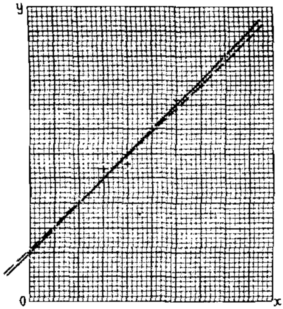


FIG. 47.—Straight line fitted by averaging.

5. Method of Averaging.—A method in common use is known as the method of averaging. Substitute the coordinates of each point for x and y in the equation.

$$y = ax + b.$$

If there are n points, this will give n equations in a and b. Unless the plotted points fall exactly on a straight line, these equations will not all be consistent. Divide the n equations into two groups with as nearly the same number of equations in each group as possible. Set up two equations in a and b such that the coefficient of a in the first equation is the arithmetic mean of the coefficients of a in the first group, the coefficient of b the arithmetic mean of the coefficients of b, and the constant term the arithmetic mean of the constant terms. Make the coefficient of a in the second equation the arithmetic mean of the coefficients of a in the second group, the coefficient of b the arithmetic mean of the coefficients of b, and the constant term the arithmetic mean of the constant terms. Solve these two equations for a and b. Substitute these values of a and b in the original equation.

Since multiplying both sides of an equation by the same number does not destroy equality, this method may be shortened by adding the equations of the first group to obtain the first equation and adding those of the second group for the second equation.

Setting up the equation for the pairs of values of x and y given on page 187 gives

$$\begin{aligned}10 &= b\\ 15 &= 5a + b\\ 20 &= 10a + b\\ 27 &= 15a + b\\ 28 &= 20a + b\\ 35 &= 25a + b\\ 41 &= 30a + b\\ 44 &= 35a + b\\ 50 &= 40a + b\end{aligned}$$

Adding the first five equations for the one and the last four for the other equation gives

$$\begin{aligned}100 &= 50a + 5b\\ 170 &= 130a + 4b.\end{aligned}$$

Solving for a and b gives

$$a = 1, b = 10.$$

The equation sought is, then,

$$y = x + 10.$$

This is a straight line with slope equal to 1 and y-intercept equal to 10. It is the full line drawn in Fig. 47.

A different grouping of the equations may give a somewhat different result. This is a defect of the method. For example, take the same set of equations and let the first group consist of the first four equations and the second group the last five. The resulting equations to solve for a and b are

$$\begin{aligned}72 &= 30a + 4b\\ 198 &= 150a + 5b.\end{aligned}$$

The resulting values of a and b are

$$a = 0.96, b = 10.8.$$

This gives

$$y = 0.96x + 10.8$$

for the required line. This has not quite as much slope as the first line found, while its y-intercept is greater. It is the dotted line shown in Fig. 47.

If three constants are required, as for the parabola

$$y = ax^2 + bx + c,$$

the equations may be divided into three groups and the same method followed. Of course, there is the same defect in the method as shown above for the straight line.

6. Least Squares Line.—A third method is by means of *least squares.* Let the equation of the required line be

$$y = ax + b,$$

in which the slope a and the y-intercept b are determined from the observed pairs of values of x and y. Take one of the given pairs of values, say, x_i and y_i. These numbers are coordinates of the point (x_i, y_i). If this point does not happen to be exactly on the required line, then $ax_i + b$ does not exactly equal y_i, but will equal the ordinate y of the point on the line whose abscissa is x_i. There is a difference, then, between the observed ordinate y_i, and the adjusted ordinate y.

$$y_i - y = y_i - (ax_i + b).$$

There is such a difference for each pair of observed values of x and y. The difference will be zero for those points, and only those points, which fall on the adjusted line.

7. The Principle Involved.—The principle of *least squares* states that the line of best fit is such a line as to *make the sum of the squares of these differences a minimum.* In other words, *the sum of the squares of the deviations of observed values of y from the corresponding values of y for the line of best fit must be less than the sum of the squares of the deviations from the corresponding values of y for any other line.*

8. The Principle Applied.—For example, take the points B_1, B_2, B_3, and B_4 in Fig. 48, plotted from a set of observations, $A_1B_1 = y_1$, $A_2B_2 = y_2$, etc. Let the line MN have the equation

$$y = ax + b.$$

The corresponding values of y are A_1C_1, A_2C_2, etc. The deviations of the ordinates of the points, B_1, B_2, etc. from the corresponding ordinates on the line MN are

$$A_1B_1 - A_1C_1 = y_1 - y = y_1 - (ax_1 + b),$$
$$A_2B_2 - A_2C_2 = y_2 - y = y_2 - (ax_2 + b).$$

The principle of least squares states that if MN is the line of closest fit, then

$$[y_1 - (ax_1 + b)]^2 + [y_2 - (ax_2 + b)]^2 + [y_3 - (ax_3 + b)]^2 + [y_4 - (ax_4 + b)]^2$$

must be less than the corresponding sum of squares of deviations from any other line. The problem is to select the parameters a and b so as to make the sum of the squares of the deviations a minimum. Represent the sum of the squares by the letter u.

The principles of differential calculus state that to make u a maximum or a minimum the derivative of u with respect to a must be zero and the derivative of u with respect to b must be zero. So, form these two derivatives, equate each to zero and solve for the unknowns, a and b.[1]

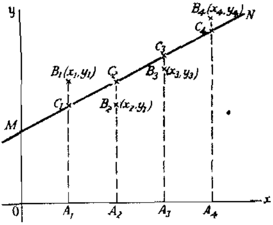


FIG. 48.—Least squares line.

Since the differences are each squared,

$$[y_1 - (ax_1 + b)]^2 = (ax_1 + b - y_1)^2.$$

Merely for convenience in writing they will be written in the latter form. Assume there are N points for which the line is to be determined.

Then

$$u = (ax_1 + b - y_1)^2 + (ax_2 + b - y_2)^2 + \cdots + (ax_N + b - y_N)^2$$

$$= \sum_{i=1}^{N}(ax_i + b - y_i)^2.$$

The derivative of u with respect to b is

$$\frac{\delta u}{\delta b} = 2(ax_1 + b - y_1) + 2(ax_2 + b - y_2) + \cdots + 2(ax_N + b - y_N)$$

$$= 2\sum_{i=1}^{N}(ax_i + b - y_i).$$

The derivative of u with respect to a is

$$\frac{\delta u}{\delta a} = 2(ax_1 + b - y_1)x_1 + 2(ax_2 + b - y_2)x_2 + \cdots + 2(ax_N + b - y_N)x_N$$

$$= 2\sum_{i=1}^{N}x_i(ax_i + b - y_i).$$

[1] See Appendix D.

Equating these derivatives to zero gives

$$2\sum_{i=1}^{N}(ax_i + b - y_i) = 0,$$

$$2\sum_{i=1}^{N}x_i(ax_i + b - y_i) = 0.$$

It is sufficient to state here that the conditions that u should be a minimum and not a maximum are also satisfied. Dividing both equations by two, expanding, and dropping the indicated limits on Σ,

$$a\Sigma x_i + Nb - \Sigma y_i = 0,$$
$$a\Sigma x_i^2 + b\Sigma x_i - \Sigma x_i y_i = 0.$$

These equations are known as *normal equations.* Systematic schemes for writing and solving general normal equations are developed in the Theory of Least Squares.

Solving for a and b and dropping the subscripts:

$$a = \frac{N\Sigma xy - \Sigma x \cdot \Sigma y}{N\Sigma x^2 - (\Sigma x)^2},$$

$$b = \frac{\Sigma x^2 \cdot \Sigma y - \Sigma x \cdot \Sigma xy}{N\Sigma x^2 - (\Sigma x)^2}.$$

Four different summations, Σx, Σy, Σx^2, and Σxy, are required.

9. Example.—Applying these results to the example just used, tabulate as in Table LIV.

TABLE LIV.—COMPUTATIONS FOR VALUES OF PARAMETERS a AND b IN THE EQUATION OF STRAIGHT LINE OF CLOSEST FIT

$N = 9$

x	y	x^2	xy
0	10	0	0
5	15	25	75
10	20	100	200
15	27	225	405
20	28	400	560
25	35	625	875
30	41	900	1,230
35	44	1,225	1,540
40	50	1,600	2,000
180	270	5,100	6,885

Substituting in the formulas for a and b gives,

$$a = \frac{9 \times 6{,}885 - 180 \times 270}{9 \times 5{,}100 - (180)^2},$$

$$b = \frac{5{,}100 \times 270 - 180 \times 6{,}885}{9 \times 5{,}100 - (180)^2}.$$

Whence

$$a = 0.99$$
$$b = 10.2.$$

The equation of the required straight line of closest fit by the method of least squares is, then,

$$y = 0.99x + 10.2.$$

10. Comparison of Results.—The slope of this line

$$a = 0.99$$

ies between the two slopes that were obtained from the two groups of equations in the method of averaging. The y-intercept,

$$b = 10.2,$$

also *lies between the two y-intercepts obtained by the method of* averaging. Collecting the values of a and b that have been obtained,

Averaging, first grouping,

$$a = 1.00,\ b = 10.00,$$

Averaging, second grouping,

$$a = 0.96,\ b = 10.8.$$

Least squares line,

$$a = 0.99,\ b = 10.2.$$

It is seen that the line of the first grouping is closer to the least squares line than is the line of the second grouping. There are no means at hand of knowing this beforehand. The labor of computation for the least squares line is not very much greater than is that for the method of averaging. Especially is this true if a computing machine or mathematical tables are available.

11. Simplification.—*It will now be shown how the formulas for* a and b, and hence the computations, may be simplified.

Take the origin of measurement of x at the arithmetic mean of x,

$$a_x = \frac{\Sigma x}{N}.$$

This has the effect of moving the y-axis to the value a_x on the x-axis (see Fig. 49). Call the new axes x'-axis and y'-axis with origin at O'. The values of the independent variable now become deviations from the arithmetic mean. Call them x'. The values of the dependent variable y remain unchanged. The formulas for a and b now are:

$$a = \frac{N \cdot \Sigma x'y - \Sigma x' \cdot \Sigma y}{N\ \Sigma x'^2 - (\Sigma x')^2},$$

$$b = \frac{\Sigma x'^2 \cdot \Sigma y - \Sigma x' \cdot \Sigma x'y}{N\ \Sigma x'^2 - (\Sigma x')^2}$$

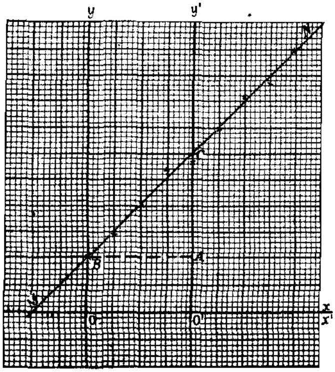


FIG. 49.—Least squares line, changed axes.

But since x' is deviation from the arithmetic mean of x

$$\Sigma x' = 0.$$

This reduces the formulas to:

$$a = \frac{N\ \Sigma x'y}{N\ \Sigma x'^2} = \frac{\Sigma x'y}{\Sigma x'^2},$$

$$b = \frac{\Sigma x'^2 \cdot \Sigma y}{N\ \Sigma x'^2} = \frac{\Sigma y}{N} = a_y.$$

The direction of the least squares line has, of course, not been changed. The y'-intercept becomes the arithmetic mean of y:

12. Example.—Applying these results to the same data as efore, Table LV results.

TABLE LV

x	y	x'	x'^2	$x'y$
0	10	−20	400	− 200
5	15	−15	225	− 225
10	20	−10	100	− 200
15	27	− 5	25	− 135
20	28	0	0	0
25	35	5	25	175
30	41	10	100	410
35	44	15	225	660
40	50	20	400	1,000
180	270	. . . . .	1,500	1,485

$$a_x = {}^{180}\!/_9 = 20. \qquad a_y = {}^{270}\!/_9 = 30.$$

$$a = \frac{1{,}485}{1{,}500} = 0.99.$$

$$b = a_y = 30.$$

In Fig. 49 the points (x, y) are plotted from the x-axis and the axis. From the values $a = 0.99$ and $b = 30$, the line MN is otted from the x'-axis and the y'-axis. Its equation referred to ese axes is

$$y' = 0.99x' + 30.$$

The value, $b = OB$, of the y-intercept, on the y-axis, may be easured graphically. It may be computed as follows:

$$\frac{AC}{BA} = a.$$

$$AC = O'C - O'A = O'C - OB = a_y - b.$$

$$BA = a_x.$$

Therefore:

$$\frac{a_y - b}{a_x} = a.$$

Solving for b gives:

$$b = a_y - a \cdot a_x.$$

In the example, then,

$$b = 30 - 0.99 \times 20,$$

$$b = 30 - 19.8,$$

$$b = 10.2.$$

So the equation, referred to x-axis and y-axis, is

$$y = 0.99x + 10.2.$$

13. Graduation or Adjustment of Observed Values.—From the equation of the line, the value of y may be obtained for any given value of x on the assumption that this least squares line is the best representation of the law connecting x and y. Also the adjustments to be made in the original observed values of y to fit this law may be computed. Computations and tabulation for the adjustment are shown in Table LVI.

TABLE LVI.—ADJUSTMENT OF OBSERVATIONS TO LEAST SQUARES LINE

Observed x	Observed y	$y = 0.99x + 10.2$	Correction to observed y
0	10	10.20	+0.20
5	15	15.15	+0.15
10	20	20.10	+0.10
15	27	25.05	−1.95
20	28	30.00	+2.00
25	35	34.95	−0.05
30	41	39.90	−1.10
35	44	44.85	+0.85
40	50	49.80	−0.20
			3.30–3.30

The observed y's all have the same weight and the least squares line passes through the average y, so that the totals of positive and negative corrections balance each other.

Applying the corrections to the observed values of y, the observations are said to be graduated to the least squares line.

14. Parabola of Closest Fit.—After plotting points for a set of observed pairs of values of two variables, x and y, judgment and experience may lead to the belief that a parabola

$$y = ax^2 + bx + c$$

could be found that would fit better than a straight line. Again, by the principle of least squares

$$u = \sum_{i=1}^{N}[ax_i^2 + bx_i + c - y_i]^2$$

must be made a minimum by proper choice of the parameters a, b, and c. So,

$$\frac{\delta u}{\delta c} = 0,$$
$$\frac{\delta u}{\delta b} = 0,$$
$$\frac{\delta u}{\delta a} = 0,$$

must be satisfied. And the result after dividing each equation by 2 and dropping subscripts is

$$\Sigma(ax^2 + bx + c - y) = 0,$$
$$\Sigma x(ax^2 + bx + c - y) = 0,$$
$$\Sigma x^2(ax^2 + bx + c - y) = 0.$$

Expanding:

$$a\Sigma x^2 + b\Sigma x + Nc = \Sigma y,$$
$$a\Sigma x^3 + b\Sigma x^2 + c\Sigma x = \Sigma xy,$$
$$a\Sigma x^4 + b\Sigma x^3 + c\Sigma x^2 = \Sigma x^2 y.$$

These are the normal equations whose solution gives the required values of a, b, and c to use in the parabola

$$y = ax^2 + bx + c$$

which best fits the set of points plotted from the observed data.

In as simple a case as the parabola, these normal equations may be written out and solved without deriving general formulas as was done for the straight line.

15. Example.—Take, for example, the following pairs of values of x and y.

x	0	5	10	15	20	25
y	10	5	10	20	35	50

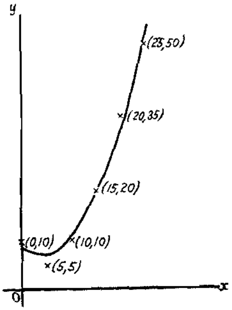


FIG. 50.—Parabola of closest fit.

Plotting the corresponding points in Fig. 50, they look as though a parabola might fit them fairly well. Now form Table LVII, making a column for each summation required by the normal equations. The first seven columns are required.

TABLE LVII.—COMPUTATION FOR VALUES OF PARAMETERS a, b, AND c, IN THE PARABOLA OF CLOSEST FIT

$$y = ax^2 + bx + c$$

x	y	x^2	x^3	x^4	xy	x^2y	y of parabola	Correction
0	10	0	0	0	0	0	8.57	−1.43
5	5	25	125	625	25	125	7.14	2.14
10	10	100	1,000	10,000	100	1,000	10.71	0.71
15	20	225	3,375	50,625	300	4,500	19.29	−0.71
20	35	400	8,000	160,000	700	14,000	32.86	−2.14
25	50	625	15,625	390,625	1,250	31,250	51.43	1.43
75	130	1,375	28,125	611,875	2,375	50,875	.....	0.00

Write the normal equations:

$$1{,}375a + 75b + 6c = 130.$$
$$28{,}125a + 1{,}375b + 75c = 2{,}375.$$
$$611{,}875a + 28{,}125b + 1{,}375c = 50{,}875.$$

Solving these equations for a, b, and c gives

$$a = 0.1000.$$
$$b = -0.7857.$$
$$c = 8.5714.$$

The required equation may now be written:

$$y = 0.10x^2 - 0.79x + 8.57.$$

16. Graduation.—Substitute the observed values of x in this equation and get the corresponding values of y for this parabola. The results are tabulated in next to the last column of Table LVII. The necessary corrections to apply to the observed values of y to graduate the data to this curve are tabulated in the last column.

17. Resulting Curve.—In Fig. 50 the values of y for the curve have been plotted and the curve drawn. It can be seen how nearly the parabola fits the observed data. A few of the points are not very close to the curve. Perhaps a curve of some other type could be found to fit the observed data better. If so, its equation would better formulate the law that would seem to govern the data. In these illustrative problems, there are not a

sufficient number of observed points to make the determination of a law very conclusive.

18. Simplification.—The y-axis may be made to pass through the average x as was done for the straight line. When this is done and abscissas become deviations from the average, $a_x = 12.5$, designated by x',

$$\Sigma x' = 0.$$

Also

$$\Sigma x'^3 = 0,$$

provided the x values are proceeding by constant differences as in the illustrative table. This simplifies the equations somewhat.

19. Other Type Forms.—It is not so simple a matter to write normal equations for graduation to fit a curve of type

$$y = ax^n$$

or

$$y = ke^{nx}.$$

Take logarithms of both sides of these equations resulting in linear equations. For the first one

$$\log y = n \log x + \log a.$$

Put

$$\log y = Y,$$
$$\log x = X,$$
$$\log a = b,$$

and the equation takes the form

$$Y = nX + b,$$

which is linear in X and Y. By tabulating the values of $\log x = X$ and $\log y = Y$, the straight line of closest fit may be obtained, by determining the values of n and b such that

$$\Sigma(nX + b - Y)^2$$

shall be a minimum. The slope n, of the line so found, is the required n of the equation of the curve. The number of which b, the Y-intercept, is the logarithm is the required a of the equation of the curve.

In the same manner, for fitting a curve of type

$$y = ke^{nx},$$

by taking logarithms of both sides, the following is obtained:

Put

$$\log y = Y,$$
$$n \log e = a,$$
$$\log k = b,$$

and the equation takes the form

$$Y = ax + b.$$

This is linear in x and Y. By tabulating values of $Y = \log y$, and plotting x with Y, a straight line of closest fit may be found. Its slope a, divided by $\log e = 0.43429$, gives the n required for the equation of the curve. The number of which the Y-intercept b is the logarithm is the k required in the equation of the curve.

20. Graphical Solution.—With paper ruled with logarithmic scale one way and uniform scale the other way, the observed values of x and y may be plotted, the straight line be drawn, and the values of n and k be determined graphically. This matter will be taken up later in the chapter on logarithmic representation.

The curve

$$y = ax^n$$

may be similarly treated on paper having logarithmic scale both ways.

21. Frequency Distributions.—One of the important problems in curve fitting is that of graduating a frequency distribution to a normal probability curve or some other recognized form whose equation is known. Professor Karl Pearson has devised certain type forms to fit almost any ordinary skew distribution.

It was seen that freshman heights (Chap. III, p. 86) fitted fairly close to the symmetric point binomial curve that resulted from the expansion of

$$(\tfrac{1}{2} + \tfrac{1}{2})^{11}.$$

22.—It is now desired to graduate the distribution of freshman heights to a normal probability curve. The equation of this curve is recalled as being

$$y = y_0 e^{-\frac{x^2}{2\sigma^2}}.$$

The function $y_0 e^{-\frac{x^2}{2\sigma^2}}$ was given (Chap. III, p. 82) as the limit of $(\tfrac{1}{2} + \tfrac{1}{2})^n$ as n increases without limit. The origin of coordinates is at the arithmetic mean. Abscissas x are deviations from the

arithmetic mean. In order that the curve may be independent of the unit measure used in measuring heights, it is plotted to the measure x/σ. The area under the curve between any two ordinates y_1 and y_2 at deviations from the arithmetic mean of x_1 and x_2, multiplied by total frequency N, gives the theoretical frequency between x_1 and x_2. The resulting theoretical frequency distribution is that called for by the laws of probability in a sample of N items selected from the universe from which the given sample was selected at random, provided the frequency distribution in that universe is a normal probability distribution. It is assumed that a_x is the arithmetic mean, and σ_x is the standard deviation of all the items in that universe.

23. The Normal Probability Curve.—It will be well to recall the shape of this curve. It is a symmetrical, bell-shaped curve.

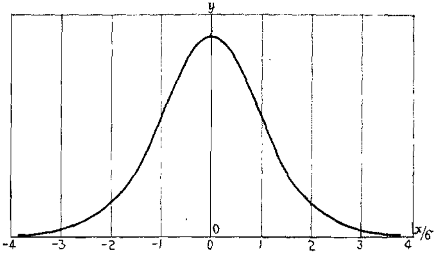


FIG. 51.—Normal probability curve.

The y-axis goes through the high point of the curve. The curve is concave downward to the points where $x/\sigma = -1$ and $x/\sigma = +1$. Beyond these points it is concave upward, getting closer and closer to the x-axis, though it never quite reaches it. Its form is shown in Fig. 51.

Tables for values of the argument x/σ have been computed, giving ordinates to this curve and areas under the curve from the y-axis to the ordinate at x/σ. In the tables the total area under the curve is taken as 1. This is assuming the total frequency to be 1; so that, for any sample of N items, the actual area is N times the tabulated area. The ordinate y_0, at $x/\sigma = 0$, has been

found to be

$$y_0 = \frac{N}{\sigma\sqrt{2\pi}};$$

so that the equation of the curve takes the form

$$y = \frac{N}{\sigma\sqrt{2\pi}}e^{-\frac{x^2}{2\sigma^2}}.$$

In order to make the measure of ordinates correspond to the x/σ measure of abscissas, and to give a total area of 1 under the curve, the actual ordinates measuring frequencies are divided by N/σ and z is put for $y \div N/\sigma$.

The equation is then of the form

$$y = \frac{N}{\sigma}z,$$

in which

$$z = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2\sigma^2}}.$$

The ordinates given in the tables are values of z.

Table LVIII is extracted from a table of ordinates and areas.

TABLE LVIII.—AREAS AND ORDINATES OF THE NORMAL PROBABILITY CURVE

$\frac{x}{\sigma}$	Area from the y-axis to the ordinate at $\frac{x}{\sigma}$	Ordinate z	$\frac{x}{\sigma}$	Area from the y-axis to the ordinate at $\frac{x}{\sigma}$	Ordinate z
0.0	0.000,00	0.3989	1.8	0.464,07	0.0790
0.1	0.039,83	0.3970	1.9	0.471,28	0.0656
0.2	0.079,26	0.3910	2.0	0.477,25	0.0540
0.3	0.117,91	0.3814	2.1	0.482,14	0.0440
0.4	0.155,42	0.3683	2.2	0.486,10	0.0355
0.5	0.191,46	0.3521	2.3	0.489,28	0.0283
0.6	0.225,75	0.3332	2.4	0.491,80	0.0224
0.7	0.258,04	0.3123	2.5	0.493,79	0.0175
0.8	0.288,14	0.2897	2.6	0.495,34	0.0136
0.9	0.315,94	0.2661	2.7	0.496,53	0.0104
1.0	0.341,34	0.2420	2.8	0.497,44	0.0079
1.1	0.364,33	0.2179	2.9	0.498,13	0.0060
1.2	0.384,93	0.1942	3.0	0.498,65	0.0044
1.3	0.403,20	0.1714	3.1	0.499,03	0.0033
1.4	0.419,24	0.1497	3.2	0.499,31	0.0024
1.5	0.433,19	0.1295	3.3	0.499,52	0.0017
1.6	0.445,20	0.1109	3.4	0.499,66	0.0012
1.7	0.455,43	0.0940	3.5	0.499,77	0.0009

It may be added that at

$$\frac{x}{\sigma} = 4.00,\ \text{Area} = 0.499{,}968{,}3,\quad z = 0.000{,}133{,}8,$$

$$\frac{x}{\sigma} = 5.00,\ \text{Area} = 0.499{,}999{,}71,\ z = 0.000{,}001{,}486{,}7,$$

$$\frac{x}{\sigma} = 6.00,\ \text{Area} = 0.499{,}999{,}99,\ z = 0.000{,}000{,}006{,}1.$$

24. Relation between σ and *p.e.*—Incidentally, it may be noted that one-fourth the area falls between $x/\sigma = 0.6$ and $x/\sigma = 0.7$. By direct interpolation in this table one-fourth of the area lies approximately at

$$\frac{x}{\sigma} = 0.67;$$

so half the total frequency lies between $x = -0.67\sigma$ and $x = +0.67\sigma$, and probable error, or quartile deviation from the median, equals 0.67σ.

25. Probability of Deviations Greater than 4σ.—At $x/\sigma = 4$ there is to the right of the y-axis 0.4999683 of the total area. And $0.50000 - 0.4999683 = 0.0000317$ of the total area lies beyond the ordinate at $x/\sigma = 4$. This makes it clear that the chances of selecting at random an item having a deviation from the arithmetic mean of more than four times the standard deviation are very small. The odds against it are 4,999,683 : 317, or about 16,000 : 1. This means that if a man's height is measured and is found to differ from the arithmetic mean of freshman heights by more than four times the standard deviation, that is, by more than

$$4 \times 2.51 \text{ in.} = 10.04 \text{ in.},$$

there is a very high probability that this man does not belong to the universe of freshmen from which the sample was taken. The arithmetic mean height was found to be 68.18 in.

$$68.18 \text{ in.} + 10.04 \text{ in.} = 78.22 \text{ in.},$$
$$68.18 \text{ in.} - 10.04 \text{ in.} = 58.14 \text{ in.}$$

By inspection of the table of freshman heights, it is found that there was not one of the 629 so tall as 78.22 in. nor so short as 58.14 in. Moreover, the chances of finding a freshman as tall

as 78 in. or as short as 58 in. are about 1 in 16,000, if the distribution of freshman heights, in general, is close to the normal probability distribution.

26. Graduation.—Now return to the problem of graduating the heights of the 629 freshmen tabulated in Table XIV to a normal frequency distribution. The necessary data and computations are tabulated in Table LIX. The first column is class boundaries.

TABLE LIX.—GRADUATION OF FRESHMAN HEIGHTS TO NORMAL PROBABILITY DISTRIBUTION

Class boundaries	x	$\frac{x}{\sigma}$	Area from y-axis to $\frac{x}{\sigma}$, A	Increment of area ΔA	Theoretical frequency $N\Delta A$	Observed frequency f	Correction to f
59.75	−8.43	−3.36	0.499,61				
				0.002,50	2	3	− 1
61.25	−6.93	−2.76	0.497,11				
				0.012,50	8	8	0
62.75	−5.43	−2.16	0.484,61				
				0.043,99	28	33	− 5
64.25	−3.93	−1.56	0.440,62				
				0.106,64	67	51	+16
65.75	−2.43	−0.97	0.333,98				
				0.189,67	119	115	+ 4
67.25	−0.93	−0.37	0.144,31				
68.18	0.00	0.00	0.000,00	0.235,26	148	156	− 8
68.75	+0.57	+0.23	0.090,95				
				0.202,94	128	148	−20
70.25	+2.07	+0.82	0.293,89				
				0.128,31	81	64	+17
71.75	+3.57	+1.42	0.422,20				
				0.056,11	35	43	− 8
73.25	+5.07	+2.02	0.478,31				
				0.017,29	11	5	+ 6
74.75	+6.57	+2.62	0.495,60				
				0.003,78	2	3	− 1
76.25	+8.07	+3.23	0.499,38				
					629	629	

$\sigma = 2.51$, $N = 629$, $a = 68.18$.

The second column is the corresponding x, or deviation from arithmetic mean. The third column is x/σ. The fourth column is A, the area from the y-axis to the ordinate at x/σ. It is obtained from the table of areas. The fifth column is ΔA, the increment of area between successive class boundaries. It is the fraction of total frequency in each of the successive classes. It is found by subtracting each entry in the A column from the

preceding entry until the class is reached in which the arithmetic mean lies. In that class ΔA equals the sum of the areas on each side of the arithmetic mean.

$$\Delta A = 0.23526 = 0.14431 + 0.09095.$$

Beyond this class ΔA for any class is the area at the upper boundary of the class minus the area at the lower boundary.

The sixth column is the theoretical frequency for each of the successive classes on the basis of total frequency equals $N = 629$ It is obtained by multiplying the entries of the fifth column by 629.

The seventh column is the observed frequency of the actual sample measured. It is the frequency column of Table XIV.

The eighth column is the correction to be applied to the observed frequency to give the theoretical frequency. It is obtained by subtracting the observed frequency from the theoretical frequency.

The curve may be plotted by plotting the theoretical frequencies at the midpoints of the classes.

27. Resulting Curve.—The resulting curve is shown in Fig. 52 If it is desired to put in intermediate points, they may be found

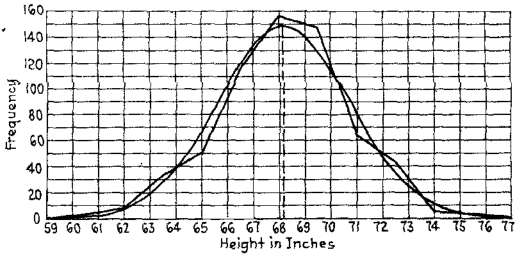


FIG. 52.—Graduation of freshman heights.

by means of Table III in Davenport's "Statistical Methods." The frequency polygon of Fig. 31 (page 86, Chap. III) is also drawn in this figure to show to the eye how closely it fits the normal probability curve.

28. Skewed Curves.—The normal probability curve of best fit has been found. If the given distribution is skewed, some other curve might fit better than the normal probability curve. The problem of fitting curves to skewed distributions is not for a beginning course in statistics.

There may be a certain amount of skewness and yet it may be permissible to graduate the data to a normal probability curve. If the measure of skewness is less than twice its probable error, it is permissible to fit a normal probability curve, provided the distribution approximates the normal type.

The skewness is then not significant. The departure from a normal probability distribution may easily be due to the accidents of sampling. If it is more than four times its probable error, it is not safe to fit a normal probability curve.

Karl Pearson has derived equations for a set of curves by means of which very many distributions may be fitted.

29. Effect of the Size of σ.—Since half the area of the normal probability curve is included between $x = -0.67\sigma$ and $x = +0.67\sigma$, it is evident that a large value of σ will make the curve spread out rather flat, while a small value of σ will make the curve narrower and run up steeper, other things being equal. The first case shows wide distribution, great variation. The latter case shows small variation, great tendency to concentrate about the mean.

30. Kurtosis.—Whether σ be large or small, it may be found that the curve of a given symmetric distribution may be flatter than the normal probability curve, may coincide with it, or may be more peaked. This shape characteristic is important and is called *kurtosis*. If the given curve is flatter than the normal probability curve, it is said to be *platykurtic*. If it is more peaked than the normal probability curve, it is called *leptokurtic*. The normal probability curve itself is said to be *mesokurtic*.

Curve fitting is also effected by what is known as the method of *moments*. This method cannot be taken up here.

TABLE LX.—MAXIMUM DAILY TEMPERATURE AT SEATTLE, JAN. 1, 1923, TO FEB. 14, 1923, INCLUSIVE, WITH MOVING AVERAGE FOR EIGHT-DAY PERIODS

Date 1923	Maximum temperature	Moving average	Date 1923	Maximum temperature	Moving average
		..			40
Jan. 1	46		24	42	
		..			40
2	49		25	41	
		..			40
3	50		26	39	
		..			39
4	46		27	40	
		50			38
5	52		28	39	
		51			38
6	53		29	32	
		50			37
7	53		30	35	
		50			37
8	51		31	38	
		49			38
9	51		Feb. 1	37	
		48			38
10	47		2	39	
		47			39
11	44		3	39	
		47			40
12	41		4	42	
		47			41
13	43		5	43	
		47			41
14	47		6	39	
		46			42
15	50		7	42	
		46			41
16	53		8	47	
		46			39
17	48		9	40	
		45			37
18	42		10	40	
		44			36
19	42		11	35	
		43			
20	40		12	29	
		42			
21	36		13	20	
		41			
22	40		14	32	
		41			
23	45				

31. Moving Average.—Sometimes a variable changes with alternate increases and decreases, in approximately periodic waves. In such case a curve of general trend may be fitted by

the method of *moving averages.* An example will best illustrate the method.

In Table LX are tabulated maximum temperatures from Jan. 1, 1923, to Feb. 14, 1923, inclusive.[1] The data are plotted to scale as shown in Fig. 53 and the points connected by straight lines from each to the next. It is noticeable that the graph fluctuates up and down in more or less regular waves. There are low points at Jan. 4, 12, 21, 29, and Feb. 6 and 13. From each to the next gives periods of 8, 9, 8, 8, and 7 days, respectively. These periods are approximately equal, with an average of 8 days. To get the moving average, find the average maximum temperature for the first 8 days. Enter this in the moving aver-

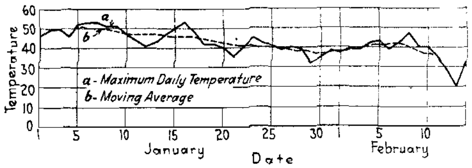


FIG. 53.—Moving average of maximum daily temperature.

age column at the middle of this 8-day period, between Jan. 4 and 5. The next entry in the moving average column is the average maximum temperature for the 8 days from Jan. 2 to 9 inclusive. The next is the average for Jan 3 to 10 inclusive. In this manner the moving average values are recorded to the last entry which is the average maximum temperature for the last 8 days of the record. A convenient way to get the values after the first is as follows: The sum of the first 8 temperatures is 400, giving an average of 50. By dropping off 46, the temperature for Jan. 1, and adding on 51, the temperature for Jan. 9, the total is increased by 5, making it 405, with an average of 51. Then by dropping off 49, the temperature for Jan. 2 and adding on 47, the temperature for Jan. 10, the total is diminished by 2, making it 403, with an average of 50. The next change in the total is $44-50$, or -6, making the third total 397, with an average of 50. The next step diminishes the total by 5, making it 392, with an average of 49. Continue the process to the end.

[1] From "Annual Meteorological Summary of the United States Weather Bureau," at Seattle, Wash.

32. Moving Average Graph or Line of Trend.—Having completed the moving average column, plot the values on the temperature graph at the dates to which they belong. Connect the points by straight lines from each to the next, or by a smooth curve. This curve shows the *general trend* of maximum temperature throughout the period except for the first four days and the last four days. If the trend line is nearly straight or changing steadily, it may usually be extended to cover the entire period.

33. What the Trend Shows.—A study of the trend line shows that there was a general tendency slowly downward through the month of January. Then there was a slight upward tendency. The daily temperatures were fluctuating back and forth on both sides of the trend. About Feb. 1 there was an upward tendency for about a week. The rest of the record shows a sharp decline. It would hardly be fair to extend this line of trend to the end of the period because the sharp decline may be due to a temporary cold spell. In fact, the sharp rise in daily temperature from Feb. 13 to 14 seems to indicate as much.

34. Short-time Fluctuations.—The deviations of daily maximum from the trend line may be plotted and thus show the short-time fluctuations independent of the long-time changes. If these deviations are to be taken from the table, it is more convenient if a moving average period with an odd number of days in it can be selected so that the entries in the moving average column will fall on exact dates. The deviations are plotted from a zero base line, above it if the deviations are positive, and below it if the deviations are negative.

Exercises

1. A rubber band, stretched under a force of p lb., is found to increase in length l mm. The following observations were made:

p..........	2	3	4	5	6	7	8	9	10	11	12
l..........	10	17	21	28	36	43	51	60	70	83	98

Find, by the method of averaging, the straight line of closest fit.

2. For the data of Ex. 1, find the straight line of closest fit by the method of least squares.

3. From the same set of data, compute and tabulate adjustments to the observed values of l to fit the least squares line of closest fit.

4. For the same set of data, by the principle of least squares, find the parabola of closest fit having the equation $l = ap^2 + bp + c$.

5. Graduate the values of l to fit the parabola obtained in Ex. 4.

6. For the same set of data in Ex. 1, find the curve of closest fit having its equation of the form

$$l = ap^n.$$

7. Plot the data of Ex. 1, and on the same axes plot the resulting curves of Ex. 1, 2, 4, and 6. Which of the four seems to give the most satisfactory fit?

8. Graduate the frequencies of freshman weights as shown in Table XIII, Chap. II, to a normal probability curve.

9. Plot the frequency polygon of freshman weigbts and the normal probability curve of Ex. 8 on the same set of axes. Does this curve seem to fit the data well?

10. Graduate the frequencies of heads falling in 500 random throws of seven dimes to a normal probability curve (see Table XV, Chap. II). Does tbis curve seem to fit the data well? Does it make a better fit than the curve of Ex. 8?

11. In the following table are the maximum temperatures each day for the month of July. Plot the values, determine the average wave length, compute a set of moving averages, and plot the results. What was the general trend of temperature during the month?

MAXIMUM TEMPERATURES DURING JULY

Day	Maximum temperature	Day	Maximum temperature
1	74	17	84
2	80	18	88
3	85	19	81
4	71	20	75
5	68	21	81
6	79	22	83
7	80	23	85
8	85	24	80
9	81	25	74
10	76	26	72
11	74	27	73
12	80	28	78
13	89	29	81
14	75	30	82
15	78	31	83
16	80		

12. Graduate the leaf lengths of Table XXIII to a normal probability curve. Does the curve appear to fit the data well?

13. Graduate the leaf breadths of Table XXIV to a normal probability curve. Does the curve appear to fit the data well?

14. Graduate the leaf lengths of Ex. 19, Chap. 11 to a normal probability curve. Does the curve appear to fit well?

15. Graduate the leaf breadths of Ex. 20, Chap. 11 to a normal probability curve. Does the curve appear to fit well?

16. A weight moving under a certain force was observed at the end of each second t to have passed over a distance s from the beginning. The results of the observations were:

t	1	2	3	4	5	6	7	8
s	0.3	1.1	2.2	3.8	6.4	9.0	12.0	16.0

Find the straight line of closest fit by the method of least squares.

17. Find, by the method of least squares, the parabola of closest fit to the data of Ex. 16.

18. Find, by the method of least squares, the curve of the best fit to the data of Ex. 16 whose equation is of the form $s = at^n$.

19. Graduate the observed values of s to fit the parabola of Ex. 17.

20. Plot the observed data of Ex. 16 and on the same axes plot the curves determined in Exs. 16 to 18. Which curve seems to fit the observed data best? Does it seem reasonable to assume that the law of motion of the weight is determined by the equation of this curve?

CHAPTER IX

CORRELATION, REGRESSION

An important question that arises in statistical study of variables is in regard to any probable relationship in the changes in one variable in connection with the changes in another variable. If one variable increases, is there a tendency for the other variable to increase or vice versa? Or, if one variable changes, does the change seem to have no influence on the changes of the other variable? Do fathers that are taller than the average of men tend, in the long run, to have sons taller than the average? If so, to what extent? Of, if the sons tend to revert, or regress, to average height, what is the measure of the tendency? Do industrial accidents tend to increase or decrease with reduction of working hours? Do bond prices tend to fluctuate with call rates of interest? Do freshman grades in college have any relation to high school grades?

1. Correlation.—Two variables are said to be *correlated* if, when any value of the first variable be selected, it is found that the average of the associated values of the other variable seems to depend on the size of the selected value of the first variable. If, when the selected value of the first variable is small, the average of the associated values of the other variable is small also, increasing with increase in size of the selected value of the first variable, the correlation is *direct* or *positive*. If, on the other hand, the average of the associated values of the second variable is large, decreasing with increase in size of the selected value of the first variable, correlation is *inverse* or *negative*.

2. Illustration.—To illustrate the meaning of correlation, Table LXI is made from Table XXII, page 38, Chap. II.

The first variable, X, represents leaf breadths, and the other, Y, represents leaf lengths. Let 19 be the selected value of X. The average value of the associated Y's is

$$\frac{2 \times 22 + 1 \times 27 + 1 \times 32}{2 + 1 + 1} = 25.75.$$

If 22 is the selected X, the average of the associated Y's is

$$\frac{1 \times 22 + 3 \times 27 + 3 \times 32 + 2 \times 37 + 1 \times 42}{10} = 31.5.$$

TABLE LXI.—SCATTER DIAGRAM OF LENGTHS AND BREADTHS OF 60 LEAVES

Measured in Millimeters

Midpoints of classes Y \ X	16	19	22	25	28	31	34	fy
52						1	1	2
47				2	3	1	1	7
42			1	3	5	3		12
37			2	5	4	3		14
32		1	3	4	3	2		13
27		1	3	3	1			8
22	1	2	1					4
fx	1	4	10	17	16	10	2	60

Thus an increase in the selected value of X of 3 mm., one class nterval, is accompanied by an increase of 31.50 mm. − 25.75 nm. = 5.75 mm. = 1.15 class interval, in the average Y associated with X.

The average Y associated with $X = 31$ is

$$\frac{2 \times 32 + 3 \times 37 + 3 \times 42 + 1 \times 47 + 1 \times 52}{10} = 40.$$

Here an increase in X of 12 mm. = 4 class intervals has given an increase in the average Y associated with X of 14.25 mm. = 2.85 class intervals.

Similarly, if $Y = 27$ be selected, the average X associated with it is

$$\frac{1 \times 19 + 3 \times 22 + 3 \times 25 + 1 \times 28}{8} = 23.5.$$

The average X associated with $Y = 42$ is 27.5. An increase in Y of 15 mm. = 3 class intervals, has given an increase in the average X associated with Y of 4.0 mm. = 1⅓ class intervals.

Inspection of the diagram readily shows that an increase in either leaf length or leaf breadth is everywhere accompanied with an increase in the average of the other associated with it. This means that there is direct correlation between lengths and breadths of leaves. This is shown by a glance of the eye, since the entries in the cells of the diagram tend to cluster about the diagonal that slopes up to the right.

If the entries cluster about the diagonal sloping down to the right, it shows that with increase of X the average Y associated with it decreases, and inverse correlation is indicated.

If the entries tend to follow neither diagonal nor any other definite line but are scattered around indiscriminately, no correlation is indicated.

3. Perfect Correlation and Independence.—If an equational relationship can be established between the selected value of the first variable and the average of the associated values of the other variable, correlation is *perfect*, positive or negative as the case may be. If the average of the associated values of the second variable seems to be completely uninfluenced by the size of the selected value of the first variable, the two variables are said to be *uncorrelated* or *independent*. It is immaterial which variable is called the *first* and which the *second*. So the same statements would apply to selected values of the second variable and the average of associated values of the first variable.

4. Measure.—The numerical measure of the tendency of the size of one variable apparently to affect the average size of the associated values of the other variable should be of the nature of a coefficient or ratio. For perfect direct correlation, its value would be $+1$, and for perfect inverse correlation, -1. For no correlation, it would be zero. One of the most used coefficients of correlation is one devised by Karl Pearson and is known as Pearson's coefficient of correlation.

5. Correlation Developed through Probability.—A good idea of the meaning of various degrees of correlation and the connection with probability may be gained by the results of tossing of coins, as shown in the following eight scatter diagrams, beginning with Table LXII.

Table LXII.—Number of Heads Falling in Pairs of Throws of Seven Dimes, All Seven Being Thrown Each Time

Number of heads in second throw	Number of heads in first throw: 0	1	2	3	4	5	6	7	Totals
0					1				1
1			3	3	4	2	2		14
2		2	8	10	10	6	1		37
3		4	7	17	16	9	2		55
4		4	5	14	22	5	4		54
5		3	2	7	9	9	1		31
6			1	2	2	2			7
7					1				1
Totals	0	13	26	53	65	33	10	0	200

Table LXIII.—Number of Heads Falling in Pairs of Throws of Seven Dimes. In the Second Throw One Marked Dime Remains as It Fell in the First Throw

Number of heads in second throw	Number of heads in first throw: 0	1	2	3	4	5	6	7	Totals
0			1						1
1	1	4		2		1			8
2			7	9	10	2			28
3	1	3	9	11	15	7	2		48
4		3	3	13	25	12	3	1	60
5	1	1	7	12	14	5	3		43
6				3	2	5			10
7				1		1			2
Totals	3	11	27	51	66	33	8	1	200

TABLE LXIV.—NUMBER OF HEADS FALLING IN PAIRS OF THROWS OF SEVEN DIMES. IN THE SECOND THROW TWO MARKED DIMES REMAIN AS THEY FELL IN THE FIRST THROW

Number of heads in second throw	Number of heads in first throw: 0	1	2	3	4	5	6	7	Totals
0		1	1	1					3
1		2	2	4	1				9
2		1	11	10	5	1	2		30
3		3	13	14	11	3	4		48
4		3	7	15	11	10	1	1	48
5			6	12	8	10	6	1	43
6		1	1		7	8	1		18
7						1			1
Totals.......		11	41	56	43	33	14	2	200

TABLE LXV.—NUMBER OF HEADS FALLING IN PAIRS OF THROWS OF SEVEN DIMES. IN THE SECOND THROW THREE MARKED DIMES REMAIN AS THEY FELL IN THE FIRST THROW

Number of heads in second throw	Number of heads in first throw: 0	1	2	3	4	5	6	7	Totals
0		1	1						2
1		5	1	4	2	1			13
2		4	11	11	10	2			38
3	1	3	12	20	13	7	1		57
4			5	12	12	6	6	1	42
5			3	2	13	7	7	2	34
6				1	4	4	2	1	12
7					1	1			2
Totals.......	1	13	33	50	55	28	16	4	200

TABLE LXVI.—NUMBER OF HEADS FALLING IN PAIRS OF THROWS OF SEVEN DIMES. IN THE SECOND THROW FOUR MARKED DIMES REMAIN AS THEY FELL IN THE FIRST THROW

		Number of heads in first throw								Totals
		0	1	2	3	4	5	6	7	
Number of heads in second throw	0				1					1
	1	2	2	2	3	1				10
	2		6	10	12	2	2			32
	3	1	5	7	19	20	3			55
	4			4	15	26	8	2		55
	5				7	11	11	4	1	34
	6				5	2		2		9
	7					1		2	1	4
Totals.......		3	13	23	62	63	24	10	2	200

TABLE LXVII.—NUMBER OF HEADS FALLING IN PAIRS OF THROWS OF SEVEN DIMES. IN THE SECOND THROW FIVE MARKED DIMES REMAIN AS THEY FELL IN THE FIRST THROW

		Number of heads in first throw								Totals
		0	1	2	3	4	5	6	7	
Number of heads in second throw	0		3							3
	1	1	1	5	6					13
	2		3	17	13	3				36
	3		1	7	14	9	2			33
	4			1	19	25	7	2		54
	5		•		8	19	16	7		50
	6						6	3	1	10
	7								1	1
Totals.......		1	8	30	60	56	31	12	2	200

TABLE LXVIII.—NUMBER OF HEADS FALLING IN PAIRS OF THROWS OF SEVEN DIMES. IN THE SECOND THROW SIX MARKED DIMES REMAIN AS THEY FELL IN THE FIRST THROW

Number of heads in second throw	Number of heads in first throw: 0	1	2	3	4	5	6	7	*Totals*
0	1								1
1		4	6						10
2		1	11	12					24
3			13	29	13				55
4				15	32	19			66
5					11	16	6		33
6							8	2	10
7							1		1
Totals......	1	5	30	56	56	35	15	2	200

TABLE LXIX.—NUMBER OF HEADS FALLING IN PAIRS OF THROWS OF SEVEN DIMES. IN THE SECOND THROW ALL SEVEN DIMES REMAIN AS THEY FELL IN THE FIRST THROW

Number of heads in second throw	Number of heads in first throw: 0	1	2	3	4	5	6	7	Totals
0	2								2
1		7							7
2			23						23
3				68					68
4					48				48
5						38			38
6							13		13
7								1	1
Totals.......	2	7	23	68	48	38	13	1	200

To obtain Table LXII seven dimes were shaken and tossed on a table, and the number of heads noted. Then the seven dimes were reshaken and tossed again and the number of heads noted. If, for example, there were three heads on the first toss and five heads on the second toss, a score mark was made in the cell in the column headed 3 and in row headed 5. This operation of taking

pairs of tosses of the seven dimes was repeated till 200 pairs of tosses were made and the score marks all entered in the diagram. The table was completed by replacing the score marks in each cell by the number of such marks.

It is evident that the number of heads falling on any single first toss could have no influence whatever on the number of heads falling on the second toss. In other words, the two variables, number of heads on the first toss, and number of heads on second toss, are completely independent and there is no correlation.

Now see what influence the size of X, the number of heads on first throw, has on $\overline{Y}$, the average number of heads on second throw, associated with X,

X	0	1	2	3	4	5	6	7
$\overline{Y}$	0	3.6	2.9	3.3	3.4	3.6	3.1	0

The change in $\overline{Y}$ seems to have nothing to do with the value of X, except that, at the extremes, $\overline{Y}$ seems fairly constant. It is likely that the extremes are accidents of the sample.

The entries in the diagram seem to cluster about the center of the diagram without spreading out in one direction more than another.

Table LXIII was obtained in a similar manner, except that one dime was blackened and left on the table so that only the other six were actually thrown on the second toss of each pair of tosses. If this black dime fell heads on the first toss, it remained heads on the second toss. Then at least one head on the second toss was certain. If the black dime fell tails on the first toss, it remained tails on the second toss. Then not more than six heads were possible on the second toss. Two hundred pairs of tosses were made, and the score marks entered and recorded as was done for Table LXII.

It was not possible to have seven heads the first toss and no heads the second toss, for if all seven dimes fell heads the first toss, the black dime fell heads and remained heads on the second toss, assuring at least one head on the second toss. Neither was it possible to have no heads the first toss, and seven heads the second toss. Diagonal lines are drawn through the cells in which no entry was possible to occur.

The number of heads falling in any first toss now has a slight influence on the number of heads falling on the corresponding second toss, for, whichever way the black dime falls and is

recorded for the first toss, it is recorded the same way for the second toss.

Working out the values of $\overline{Y}$ for each value of X, as was done before, the result is:

X	0	1	2	3	4	5	6	7
$\overline{Y}$	3.0	2.7	3.3	3.7	3.7	4.1	4.1	4.0

It is discovered that there is a slight tendency such that the larger the number of heads on the first toss the larger will be the number of heads on the second toss. In other words, some correlation is indicated.

Table LXIV was obtained in exactly the same manner except that two dimes were blackened and left on the table after each first toss so that the two remained unchanged on the corresponding second toss.

Now *if all seven dimes fell heads on the first toss, there would* have to be at least two heads on the second toss, for the two black dimes would be both heads. If six dimes fell heads on the first toss, there would have to be at least one head on the second toss, for at least one of the black dimes would be heads. Similarly, if there were no heads on the first toss, there could not be more than five heads on the second toss, and if there were one head on the first toss, there could not be more than six on the second toss. Diagonal lines are drawn to cut off the cells in which entries are impossible to occur.

Determination of values of $\overline{Y}$ for each X gives:

X	0	1	2	3	4	5	6	7
$\overline{Y}$	0	2.8	3.1	3.3	3.9	4.7	4.0	4.5

In only one case does $\overline{Y}$ decrease for increase of X, though for $X = 6$ and $X = 7$, $\overline{Y}$ is less than for $X = 5$.

Table LXV is obtained in exactly the same way except that three dimes are blackened and left on the table for each second toss. The diagonal lines cutting off the cells in which entries are impossible cut off more cells than before, crowding the entries closer to the diagonal that slopes down to the right. The increased *tendency to follow this diagonal indicates stronger correlation.* (Why not the diagonal up to the right? Some statisticians follow this method of tabulating the variables and some the method used in Table LXI.)

The values of $\overline{Y}$ corresponding to each X are:

X	0	1	2	3	4	5	6	7
$\overline{Y}$	3	1.7	2.8	3.0	3.7	4.1	4.6	5.0

Except for $X = 0$, $\overline{Y} = 3$, the values of $\overline{Y}$ are running along with the corresponding values of X more closely than in the previous tables. The exceptional value of $\overline{Y}$ for $X = 0$ is due to a single pair of tosses and is undoubtedly an accident of the sample.

In making Table LXVI four dimes were blackened and remained unchanged for the second toss of each pair. The diagonal lines are drawn still closer to the main diagonal of the table. The entries cluster still more closely about this main diagonal. The values of X and $\overline{Y}$ follow:

X	0	1	2	3	4	5	6	7
$\overline{Y}$	1.7	2.2	2.6	3.4	3.9	4.2	5.4	6.0

In making Table LXVII five dimes were blackened, remaining unchanged in each second toss. The entries crowd closer to the main diagonal. The values of X and $\overline{Y}$ are:

X	0	1	2	3	4	5	6	7
$\overline{Y}$	1.0	1.2	2.1	3.2	4.1	4.8	5.1	6.5

Now for nearly every value of X the value of $\overline{Y}$ is quite close to the value of X.

For Table LXVIII six dimes were blackened and remained unchanged for each second toss. All entries are now very close to the main diagonal down to the right. The corresponding values of X and $\overline{Y}$ are:

X	0	1	2	3	4	5	6	7
$\overline{Y}$	0	1.2	2.2	3.1	4.0	4.5	5.7	6.0

In getting Table LXIX all the dimes of the first toss were left on the table and counted as they were for the second toss. The entries are all confined to the main diagonal down to the right. As soon as the value of X is known, the value of $\overline{Y}$ is known to be the same. Everywhere

$$\overline{Y} = X.$$

There is an equational relationship and perfect correlation.

This survey has now gone through these eight tables recording the values of $\overline{Y}$, the average number of heads in second toss corresponding to X, the number of heads in first toss. It is found

that, by making the tosses in such a manner that the number of heads, X, in the first toss has more and more influence in determining how many heads shall fall on the second toss, $\bar{Y}$ has a stronger and stronger tendency to become equal to its corresponding X. This means that correlation is becoming greater. It is also noticed in the diagrams that the entries have a growing tendency to follow a diagonal of the figure as correlation increases. If for each table there had been 2,000 pairs of tosses of dimes instead of 200, these tendencies would undoubtedly have been more marked.

In the first table there was very little correlation, if any. From the fact that the number of heads in the second toss was absolutely independent of the number in the first toss, it must be concluded that, if any correlation is indicated, it is an accident of the sample.

One might just as well have associated the average number of heads on the first toss that occurred with each possible number on the second toss, that is, have found $\bar{X}$ for each Y. The results would have been similar. Increasing correlation would have been observed as progress was made through the successive tables, ending with the perfect correlation of Table LXIX.

6. Regression Lines and Equations.—An expression for a measure of correlation may be obtain by first deriving the equations of what are known as *regression lines* or *characteristic lines*.

Turning again to Table LXI, calculate the average Y associated with each X. This means—calculate the average length of all leaves having each given breadth. In other words, find the average of each of the columns. Let $\bar{Y}$ stand for this mean of each column. Take a pair of coordinate axes and plot the points $(X, \bar{Y})$, one for each column of the table. Find the straight line of closest fit by the method of least squares. Take the y-axis through a_X, the arithmetic mean of breadths X, and the x-axis through a_Y, the arithmetic mean of lengths. Use x for deviations from a_X, and y for deviations from a_Y. For any value of x, the items in the corresponding column give an array of values of y. This array is called the y-array of type x. Let n be the number of items in any y-array, n' being the number in the first y-array, n'' the number in the second y-array, and so on to $n^{(r)}$, the number in the rth y-array, there being r such arrays in all. Let the equation of the required line of closest fit be

$$\bar{y} = b_1x + a_1.$$

It is required to make the coefficients a_1 and b_1 such that

$$\Sigma[\bar{y} - (b_1x + a_1)]^2$$

shall be a minimum.

Let RR' (Fig. 54) be such a line. For any given $x = OA_1$,

$$b_1x + a_1 = A_1C_1$$

and

$$\bar{y} = A_1B_1.$$

$$\bar{y} - (b_1x + a_1) = C_1B_1$$

is the distance the arithmetic mean of the y-array of type x is from the line RR'. The sum of the squares of these distances, one for each of the N pairs of items in the entire table, must be a

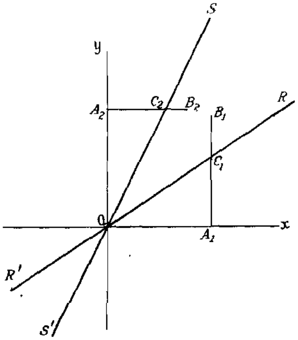


FIG. 54.

minimum. In Chap. VIII, page 192, it was determined that a_1 and b_1 must satisfy the normal equations:

$$\Sigma[(b_1x + a_1) - \bar{y}] = 0$$

and

$$\Sigma x[(b_1x + a_1) - \bar{y}] = 0.$$

Now, any y-array of type x, containing n items, contributes n terms to each of the above summations. The sum of $\bar{y}$ in the x-array will then be $n\bar{y}$, and the sum of $\bar{y}$ for the entire table will be $\Sigma n\bar{y}$, n and $\bar{y}$ each varying from column to column. The sum of $x\bar{u}$ for the entire table will be $\Sigma x(n\bar{y})$.

Expanding and transposing, the normal equations now become:

$$b_1\Sigma x + Na_1 = \Sigma n\bar{y},$$
$$b_1\Sigma x^2 + a_1\Sigma x = \Sigma x(n\bar{y}).$$

In any y-array,

$$n\bar{y} = \Sigma y$$

for the n items in that array.

Then for the entire table

$$\Sigma n\bar{y} = \Sigma y.$$

Similarly,

$$\Sigma x(n\bar{y}) = \Sigma xy$$

for the entire table.

The normal equations may now be written:

$$b_1\Sigma x + Na_1 = \Sigma y,$$

and

$$b_1\Sigma x^2 + a_1\Sigma x = \Sigma xy.$$

But since x is deviation from a_X and y is deviation from a_Y,

$$\Sigma x = 0 = \Sigma y.$$

Therefore,

$$Na_1 = 0$$

and

$$a_1 = 0.$$

Also,

$$b_1\Sigma x^2 = \Sigma xy$$

and

$$b_1 = \frac{\Sigma xy}{\Sigma x^2}.$$

Now introduce what is known as the *product moment,*

$$p = \frac{\Sigma xy}{N}$$

and a quantity r such that

$$r = \frac{p}{\sigma_X\sigma_Y}.$$

Then,

$$b_1 = \frac{Np}{N\sigma_X^2} = \frac{r\sigma_X\sigma_Y}{\sigma_X^2},$$

so that

$$b_1 = r\frac{\sigma_Y}{\sigma_X}.$$

The equation of the line of closest fit is, therefore;

$$\bar{y} = r\frac{\sigma_Y}{\sigma_X}\cdot x.$$

This is the *regression equation* of y on x. The line is the *regression line* of y on x.

$$b_1 = r\frac{\sigma_Y}{\sigma_X}$$

is called the *regression coefficient* of y on x. Applied in Table LXI, it would be the regression coefficient of leaf length on leaf breadth.

7. Meaning of Regression Line.—This regression line is a line such that the sum of the squares of the distances of the arithmetic means of the columns, from the regression line, each mean counted as many times as there are items in the column, is the least possible. If some other line were drawn, the sum of the squares of the distances from it would be more than when this regression line is used.

The equation of the line of regression may be regarded as an equation from which to estimate the average y associated with each x in such a manner as to make the sum of the squares of the errors of estimate a minimum. Or, it may be considered as that equation from which $\bar{y}$, the average of all the y's associated with any value of x, may be estimated so that the sum of the squares of the errors of estimate shall be a minimum, when each mean, $\bar{y}$, is weighted by the number of items in the y-array determining it. It is the line which best fits the points determined by the averages of the columns in the scatter diagram.

8. Another Regression Line.—In the same manner may be found the line of best fit to the points determined by the averages of the rows in the scatter diagram. Its equation would be the form

$$\bar{x} = b_2y + a_2.$$

Let SS' (Fig. 54) be such a line. For any x-array of type $y = OA_2$, let $\bar{x} = A_2B_2$ be the average x in that array, and $A_2C_2 = b_2y + a_2$.

Determine b_2 and a_2 so that

$$\Sigma[\bar{x} - (b_2y + a_2)]^2$$

shall be a minimum. A little consideration of the symmetry of the figure will show that the required regression equation is the equation of RR' with x and y everywhere interchanged. Interchange of x and y does not change r. Therefore, the required equation of SS' is

$$\bar{x} = r\frac{\sigma_X}{\sigma_Y}\cdot y.$$

$$a_2 = 0$$

and

$$b_2 = r\frac{\sigma_X}{\sigma_Y}.$$

This equation is the regression equation of x on y, and b_2 is the regression coefficient of x on y. The equation is the equation for estimating $\bar{x}$, the average x in the x-array associated with any y, in such a manner as to make the sum of the squares of the errors of estimate a minimum, each $\bar{x}$ being counted as many times as there are items in the x-array to which it belongs.

The slope of RR' (Fig. 54) equals A_1C_1/OA_1.

So

$$b_1 = r\frac{\sigma_Y}{\sigma_X} = \frac{A_1C_1}{OA_1}.$$

Similarly,

$$b_2 = r\frac{\sigma_X}{\sigma_Y} = \frac{A_2C_2}{OA_2}.$$

9. Regression Coefficients are Constants.—The two standard deviations, σ_X and σ_Y, are statistical constants which can be computed from the frequency tables. The quantity $r = p/\sigma_X\sigma_Y$, where $p = \Sigma xy/N$, may also be computed from the scatter diagram, or correlation table. Hence b_1 and b_2 are statistical constants which may be at once computed from the correlation table. With b_1 and b_2 known, the two regression equations may be written at once.

10. Averages Exactly on Straight Lines.—If the averages of the columns happen to fall exactly on a straight line RR', then, upon selecting any x determining a y-array, the average of the items in that array is known at once from

$$\bar{y} = r\frac{\sigma_Y}{\sigma_X}\cdot x,$$

the equation of RR'. If the line slopes up to the right, as in Fig. 54, then the larger the x selected the larger the average y associated with it. This would indicate direct correlation. Similarly, if the averages of the rows happen to fall exactly on a straight line SS', then the average x in any x-array of type y may be found from the regression equation

$$\bar{x} = r\frac{\sigma_X}{\sigma_Y}\cdot y.$$

If the line slopes as does SS' in Fig. 54, then the larger the y selected the larger is the average of the x's associated with it. This would indicate direct correlation.

11. Averages Not on Straight Lines.—It seldom happens that either set of averages falls exactly on a straight line. However, they frequently fall sufficiently close to straight lines so that the deviations from the regression line may be assumed as errors of sampling. If they are so scattered that this assumption is not true, then the regression line, the straight line of closest fit, has no significance. In such case, it may happen that the averages tend to follow some other mathematical curve and so still show correlation. If the averages do tend to follow a straight line, the regression is called *linear regression*. If the averages tend to follow some other curve, there is said to be *non-linear regression*, and the correlation is called *skew correlation*.

12. Coincidence of Regression Lines.—If the averages of the columns and the averages of the rows fall on the same straight line, there is equational relationship between the two sets of averages indicating perfect correlation. In that case, the lines RR' and SS' coincide.

13. Regression Line Falling along a Coordinate Axis.—If the averages of the columns fall all on the x-axis, RR' coincides with the x-axis. The size of any selected x would then have no influence on the average of the associated y-array. Similarly, if SS' coincides with the y-axis, the size of any selected y has no influence on the average of the associated x-array. In either case, there is no connection between the size of one variable and the associated sizes of the other variable. In other words, there is no correlation.

14. Regression Lines and Correlation.—Evidently, the nearer RR' and SS' come to coincidence, the greater the degree of correlation between the two variables. The nearer RR' approaches the x-axis or SS' approaches the y-axis, the slighter the degree of correlation.

15. The Product b_1b_2 as a Measure of Correlation.—If RR' and SS' coincide,

$$b_1 = \frac{1}{b_2}.$$

This is seen at once from Fig. 54. If OC_1 and OC_2 are taken equal and either line (or both) is rotated to coincide with the other, C_1 and C_2 will coincide and then,

$$OA_1 = A_2C_2$$

and

$$A_1C_1 = OA_2.$$

Then,

$$\frac{A_1C_1}{OA_1} \text{ will equal } \frac{OA_2}{A_2C_2}.$$

But

$$\frac{A_1C_1}{OA_1} = b_1 \text{ and } \frac{OA_2}{A_2C_2} = \frac{1}{b_2}.$$

Therefore,

$$b_1 = \frac{1}{b_2}$$

in this case.

Then,

$$b_1b_2 = 1.$$

If RR' coincides with the x-axis,

$$b_1 = 0.$$

If SS' coincides with the y-axis,

$$b_2 = 0.$$

In either case,

$$b_1b_2 = 0.$$

The farther RR' and SS' recede from the x-axis and the y axis, respectively, and the nearer they approach coincidence the greater the degree of correlation, and the nearer does $b_1b_2 = 1$

The quantity b_1b_2 then makes a good measure of correlation i the case of linear regression, since, when it becomes 0, there is n correlation, and the nearer it gets to 1, the greater the degree c correlation, showing perfect correlation when it equals 1.

If, when RR' and SS' coincide, the line slopes up to the right the correlation is direct or positive. If it slopes down to th right, correlation is inverse or negative.

16. The Quantity r as a Measure of Correlation.—Now

$$b_1b_2 = r\frac{\sigma_Y}{\sigma_X} \cdot r\frac{\sigma_X}{\sigma_Y}.$$

Therefore,

$$r^2 = b_1b_2$$

and

$$r = \sqrt{b_1b_2}.$$

It follows that r becomes an excellent measure of correlation in the case of linear regression. If the correlation is direct, the plus sign is used with r. If correlation is inverse, the minus sign is used.

It is necessary that b_1 and b_2 have the same sign, or r would be an imaginary quantity. In Fig. 54, b_1 and b_2 are both positive. If RR' and SS' sloped down to the right, both b_1 and b_2 would be negative. In either case, b_1b_2 is positive and r is real; r is the geometric mean of b_1 and b_2.

17. Formula for r.—Going back to the original definition of r on page 224:

$$r = \frac{p}{\sigma_X \sigma_Y}$$

and

$$p = \frac{\Sigma xy}{N}.$$

That is,

$$r = \frac{\Sigma xy}{N\sigma_X \sigma_Y}.$$

This is the form in which it is usually written. It is known as *Pearson's coefficient of correlation*. It is applicable to distributions having approximate linear regression. In case of skew correlation, some other measure of correlation must be used. Translating the formula for r into English, it reads: The coefficient of correlation equals the product moment of the associated variables divided by the product of their standard deviations.

Substituting $\Sigma x^2/N$ for $\sigma_X{}^2$ and $\Sigma y^2/N$ for σy^2 in

$$r = \frac{\Sigma xy}{N\sigma_X \sigma_Y}$$

gives

$$r = \frac{\Sigma xy}{N\sqrt{\frac{\Sigma x^2}{N} \cdot \frac{\Sigma y^2}{N}}}$$

or

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \cdot \Sigma y^2}}.$$

If the standard deviations of the variables are not needed, this formula gives a more direct means of computing r.

18. Short-cut Formula.—Let the short-cut methods of computing standard deviation and average be recalled. For a_X, the average of X, an arbitrary origin X' was chosen and it was found that

$$a_X = X' + \frac{\Sigma d_{X'}}{N},$$

the arbitrary value chosen plus the average deviation from the arbitrary value chosen.

It was also found that

$$\sigma_X = \sqrt{\frac{\Sigma d^2_{X'}}{N} - (a_X - X')^2}.$$

Let

$$a_X - X' = \bar{x}'.$$

From the short-cut equation for a_X

$$\bar{x}' = \frac{\Sigma d_{X'}}{N}.$$

Let $x' = d_{X'} = X - X'$ and $y' = d_{Y'} = Y - Y'$.

Now the short-cut equations may be written

$$a_X = X' + \bar{x}',$$

$$\sigma_X = \sqrt{\frac{\Sigma x'^2}{N} - \bar{x}'^2}.$$

Similarly,

$$a_Y = Y' + \bar{y}'^2$$

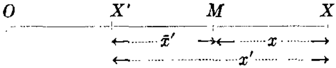
and

$$\sigma_Y = \sqrt{\frac{\Sigma y'^2}{N} - \bar{y}'^2}.$$

It is now desired to find an expression for product moment, $\Sigma xy/N$, in terms of x' and y'.

$$\begin{aligned} x' &= X - X' \\ &= (x + a_X) - (a_X - \bar{x}') \\ &= x + \bar{x}'. \end{aligned}$$

This may be seen graphically by the following diagram, in which O is the origin of measurements:



$OX' = X'$, $OM = a_X$, $OX =$ any X.
$X'M = a_X - X' =$
$MX = X - a_X = x$
$X'X = X - X' =$:

Now directly from the figure:

$$x' = x +$$

true for every X.

Similarly,

$$y' = y + \bar{y}'.$$

Then,

$$x'y' = (x + \bar{x}')(y + \bar{y}')$$
$$= xy + \bar{x}'y + \bar{y}'x + \bar{x}'\bar{y}'.$$

So

$$\Sigma x'y' = \Sigma xy + \bar{x}'\Sigma y + \bar{y}'\Sigma x + N\bar{x}'\bar{y}'.$$

But

$$\Sigma x = 0 = \Sigma y.$$

since x and y are deviations from the respective arithmetic means.

Therefore,

$$\Sigma x'y' = \Sigma xy + N\bar{x}'\bar{y}'$$

and

$$\Sigma xy = \Sigma x'y' - N\bar{x}'\bar{y}'.$$

The formula for r now becomes:

$$r = \frac{\Sigma x'y' - N\bar{x}'\bar{y}'}{N\sigma_X\sigma_Y}.$$

19. Arbitrary Origin at Zero.—As a special case, take the arbitrary numbers X' and Y' at the origin of measurements, that is:

$$X' = Y' = 0.$$

Then,

$$x' = X,\ y' = Y,$$

$$a_X = \frac{\Sigma X}{N} = \bar{x}',\ a_Y = \frac{\Sigma Y}{N} = \bar{y}',$$

$$\sigma_X = \sqrt{\frac{\Sigma X^2 - Na_X^2}{N}},\ \sigma_Y = \sqrt{\frac{\Sigma Y^2 - Na_Y^2}{N}},$$

$$\Sigma xy = \Sigma XY - Na_Xa_Y.$$

And

$$r = \frac{\Sigma XY - Na_Xa_Y}{N\sqrt{\frac{\Sigma X^2 - Na_X^2}{N}} \cdot \sqrt{\frac{\Sigma Y^2 - Na_Y^2}{N}}},$$

or

$$r = \frac{\Sigma XY - Na_Xa_Y}{\sqrt{(\Sigma X^2 - Na_X^2)(\Sigma Y^2 - Na_Y^2)}}.$$

This gives r in terms of the original measurements X and Y, and their arithmetic means.

20. Construction of a Correlation Table.—Now take Table LXI and extend it for getting the arithmetic means and standard deviations of the two variables, length and breadth, and the coef-

ficient of correlation between these two variables. This forms Table LXX.

The *arbitrary origin of measurement should be near the middle* of the distribution.

Let

$$X' = 25 \text{ mm.}$$

and

$$Y' = 37 \text{ mm.}$$

The classes having these midpoint values are ruled off by heavy lines, for convenience. *Let the class interval be the unit of measurement.*

$$X' = 25/3 = 8\tfrac{1}{3} \text{ class intervals.}$$
$$Y' = 37/5 = 7.4 \text{ class intervals.}$$

Then deviations, x' and y' from X' and Y', are multiples of class intervals. Columns are made for y', fy', fy'^2, $fx'y'$, and $f(y' + 1)^2$. Additional rows are made for x', fx', fx'^2, $fx'y'$, and $f(x' + 1)^2$.

In the y' column, begin with 0 at the arbitrary origin and write 1, 2, etc. upward to the top of the table. Write -1, -2, etc. downward to the bottom of the table, the class interval being the unit of measurement. Column fy' is simply the product of f by y'; column fy'^2, the product of y' and fy'. The rows x', fx', fx'^2 are obtained in an exactly similar manner. Column $fx'y'$ is obtained as follows: In the lower right-hand corner of the occupied cells are written small figures. Each of these is the $x'y'$ product for each item in the cell. For each cell in column headed 25, $x' = 0$. Then $x'y' = 0$ for each cell in that column. For each cell in column headed 31, $x' = 2$. For each cell in column headed 19, $x' = -2$. For each cell in any column x' equals the x' entry for that column. Similarly, the y' for each cell in any row is the y' entry for that row. Then the $x'y'$ entry for any cell is the product for the x' entry for the column in which the cell stands, by the y' entry for the row in which it stands. For column headed 28, $x' = 1$. So the small figures are each one times the corresponding y'. For column headed 31, $x' = 2$. So the small figures are each two times the corresponding y'. That is, they start at 0 and go up and down by multiples of 2 and -2, respectively. For column headed 34, the small figures proceed from 0 up and down by multiples of $+3$ and -3, respectively. Similar formation of the $x'y'$ products for every cell of

the table follows readily. If much correlation work is to be done, it may well pay to have printed sheets ruled and the proper $x'y'$ entry printed in each cell.

Having completed the proper small figure entries in the cells and having recorded frequencies, at once obtain the $fx'y'$ entries. In each cell, of course, there is an $x'y'$ for each item that falls in that cell. So multiply each cell frequency in a row by the $x'y'$ for the cell and sum across the entire row. Thus, for the first row,

$$fx'y' = 1 \times 6 + 1 \times 9 = 15.$$

For the second row,

$$fx'y' = 2 \times 0 + 3 \times 2 + 1 \times 4 + 1 \times 6 = 16.$$

For the third row,

$$fx'y' = 1 \times (-1) + 3 \times 0 + 5 \times 1 + 3 \times 2 = 10.$$

The fourth row gives 0.

For the fifth row,

$$fx'y' = 1 \times 2 + 3 \times 1 + 4 \times 0 + 3 \times (-1) + 2 \times (-2) = -2.$$

Continue the process for all rows, entering the sums in the $fx'y'$ column. Be sure to observe algebraic signs. A similar process applied to the columns gives the row headed $fx'y'$. This row is merely inserted as a check. $fx'y'$ should be the same both ways. The $f(y' + 1)^2$ column is self-explanatory. For any row, add 1 to y', square, and multiply by f. For the first row,

$$(3 + 1)^2 \times 2 = 32.$$

The row $f(x' + 1)^2$ is obtained similarly. Find the algebraic sum of every column and row except column y' and row x'.

21. Charlier Check.—The column $f(y' + 1)^2$ is the Charlier check on the other work.

$$f(y' + 1)^2 = fy'^2 + 2fy' + f,$$

So

$$\Sigma f(y' + 1)^2 = \Sigma fy'^2 + 2\Sigma fy' + N.$$

In this table,

$$\Sigma f(y' + 1)^2 = 181,$$

$$\Sigma fy'^2 + 2\Sigma fy' + N = 139 + 2(-9) + 60 = 181,$$

which checks.

Similarly, for $f(x' + 1)^2$

$$211 = 109 + 2 \times 21 + 60.$$

TABLE LXX.—CORRELATION TABLE FOR LENGTHS AND BREADTHS OF 60 LEAVES; X = BREADTHS, Y = LENGTHS, IN MILLIMETERS

Midpoints Y \ X	16	19	22	25	28	31	34	f	y'	fy'	fy'^2	$fx'y'$	$f(y'+1)^2$	$\bar{x}'_{av}$	$\bar{x}'^2_{av}$	$f\bar{x}'^2_{av}$
52						1_{6}	1_{9}	2	3	6	18	15	32	2.50	6.25	12.50
47				2_{0}	3_{2}	1_{4}	1_{6}	7	2	14	28	16	63	1.14	1.30	9.10
42			1_{-1}	3_{0}	5_{1}	3_{2}		12	1	12	12	10	48	0.83	0.69	8.28
37			2_{0}	5_{0}	4_{0}	3_{0}		14	0	0	0	0	14	0.57	0.32	4.48
32		1_{2}	3_{1}	4_{0}	3_{-1}	2_{-2}		13	−1	−13	13	−2	0	0.15	0.22	2.86
27		1_{4}	3_{2}	3_{0}	1_{-2}			8	−2	−16	32	8	8	−0.50	0.25	2.00
22	1_{9}	2_{6}	1_{3}					4	−3	−12	36	24	16	−2.00	4.00	16.00
f	1	4	10	17	16	10	2	60		−9	139	71	181			55.22
x'	−3	−2	−1	0	1	2	3									
fx'	−3	−8	−10	0	16	20	6	21								
fx'^2	9	16	10	0	16	40	18	109								
$fx'y'$	9	18	11	0	6	12	15	71								
$(x'+1)^2$	4	4	0	17	64	90	32	211								
$\bar{y}'_{av}$	−3.00	−2.25	−1.10	−0.18	0.37	0.60	2.50									
$\bar{y}'^2_{av}$	9.00	5.06	1.21	0.03	0.14	0.36	6.25									
$f\bar{y}'^2_{av}$	9.00	20.24	12.10	0.54	2.24	3.60	12.50	60.22								

22. Computation of the Constants.—Now obtain the constants used in the formulas.

$$\bar{x}' = \frac{\Sigma x'}{N} = \frac{21}{60} = 0.35.$$

$$a_X = X' + \bar{x}' = 8.33 + 0.35 = 8.68.$$

$$\sigma_X = \sqrt{\frac{\Sigma x'^2}{N} - \bar{x}'^2} = \sqrt{\frac{109}{60} - (0.35)^2} = 1.30.$$

$$\bar{y}' = \frac{\Sigma y'}{N} = \frac{-9}{60} = -0.15.$$

$$a_Y = Y' + \bar{y}' = 7.40 + (-0.15) = 7.25,$$

$$\sigma_Y = \sqrt{\frac{\Sigma y'^2}{N} - \bar{y}'^2} = \sqrt{\frac{139}{60} - (-0.15)^2} = 1.51.$$

All these results are in class intervals as units of measurement. To get the results in millimeters, the X-results, breadths, must be multiplied by 3, the X class interval, and the Y-results, lengths, by 5, the Y class interval. This gives

$$a_X = 8.68 \times 3 = 26.04 \text{ mm.}$$
$$a_Y = 7.25 \times 5 = 36.25 \text{ mm.}$$
$$\sigma_X = 1.30 \times 3 = 3.90 \text{ mm.}$$
$$\sigma_Y = 1.51 \times 5 = 7.55 \text{ mm.}$$

23. Value of r.—For Pearson's coefficient of correlation,

$$r = \frac{\Sigma x'y' - N\bar{x}'\bar{y}'}{N\sigma_X\sigma_Y} = \frac{71 - 60(0.35)(-0.15)}{60 \times 1.30 \times 1.51}.$$

Whence,

$$r = 0.63.$$

The formula for probable error of r has been given as

$$p.e. \text{ of } r = \frac{0.6745(1 - r^2)}{\sqrt{N}}.$$

In this problem

$$p.e. \text{ of } r = \frac{0.6745(1 - 0.397)}{\sqrt{60}} = 0.053.$$

So the coefficient of correlation is written

$$r = 0.63 \pm 0.05.$$

The chances are even that the true value of r differs from 0.63 by not more than 0.05.

In the formula for r, it is not necessary to multiply x' and y' by the respective class intervals. If this were done, both numerator and denominator would be multiplied by both class intervals, and the value of the fraction would not be changed.

In this problem, the numerator and the denominator would each be multiplied by $3 \times 5 = 15$.

Other measures of the degree of correlation have been devised. *Pearson's coefficient, the product moment method, is based on a* standard deviation about the line of regression, since it involves the sum of the squares of the deviations of the row (and column) averages from the line of regression.

24. Correlation Ratio.—If there is skew correlation, that is, if the row (or column) averages do not tend to follow a straight line, but some other curve, a method would be based on an average standard deviation with respect to such curve.

Let s_{ax} be the standard deviation of any x-array consisting of n items. Let σ^2_{ax} be a weighted mean of all the s^2_{ax}'s, the weights being the corresponding values of n. Then,

$$\sigma^2_{ax} = \frac{\Sigma n s^2_{ax}}{N}.$$

This σ_{ax} is based on a line going through the average of the rows, since every s_{ax} is a standard deviation from the average of the corresponding row. It can be seen that this weighted mean of standard deviations of row arrays will be less than the standard deviation of the entire set of N observations.[1]

Then write

$$\sigma^2_{ax} = \sigma_X^2(1 - \eta^2_{XY}).$$

Solving for, η^2_{XY}, (*eta* square, X on Y),

$$\eta^2_{XY} = \frac{\sigma^2_X - \sigma^2_{ax}}{\sigma_X^2}.$$

Or

$$\eta^2_{XY} = 1 - \frac{\sigma^2_{ax}}{\sigma_X^2}.$$

Thus, η^2_{XY} is the ratio of the reduction of the square of standard deviation to the square of the standard deviation itself when the weighted average square standard deviation of rows is used. Professor Pearson calls η the *correlation ratio*. It may also be shown[2] that

$$\sigma_X^2 - \sigma^2_{ax} = \sigma^2_{mx}$$

where σ_{mx} is the weighted standard deviation of the means of the x-arrays from a_X. Therefore, write

$$\eta^2_{XY} = \frac{\sigma^2_{mx}}{\sigma_X^2}$$

[1] This may be seen in an example by inspection of the table, in which the standard deviation of each row (or column) is less than that of the entire table. The mathematical proof is not given here.

[2] G. UDNY YULE, "Theory of Statistics," p. 205 (4).

or

$$\eta_{XY} = \frac{\sigma_{mx}}{\sigma_X}.$$

It is evident that if one starts with, say, standard deviation of any y-array, another value of η designated by η_{YX} will result. The same line of reasoning will give

$$\eta_{YX} = \frac{\sigma_{my}}{\sigma_Y}.$$

26. Example.—In Table LXX are placed additional columns for determining σ_{mx}, and additional rows for determining σ_{my}. Having these determined, the correlation ratio of X on Y, η_{XY}, and of Y on X, η_{YX}, may be at once computed.

In Table LXX, the column heading $\bar{x}'_{ax}$ is the deviation of the average of an x-array, measured in class intervals, from the arbitrary origin.

For the first x-array,

$$\bar{x}'_{ax} = \frac{1 \times 2 + 1 \times 3}{2} = 2.50.$$

For the second x-array,

$$\bar{x}'_{ax} = \frac{2 \times 0 + 3 \times 1 + 1 \times 2 + 1 \times 3}{7} = 1.14.$$

For the third x-array,

$$\bar{x}'_{ax} = \frac{1 \times (-1) + 3 \times 0 + 5 \times 1 + 3 \times 2}{12} = 0.83.$$

And so on through all the x-arrays. The next two columns are self-explanatory.

The row heading $\bar{y}'_{ay}$ is the deviation of the average of a y-array, measured in class intervals, from the arbitrary origin.

For the first y-array,

$$\bar{y}'_{ay} = \frac{1 \times (-3)}{1} = -3.00.$$

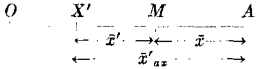
For the fourth y-array,

$$\bar{y}'_{ay} = \frac{2 \times 2 + 3 \times 1 + 5 \times 0 + 4 \times (-1) + 3 \times (-2)}{17} = -0.18.$$

And so on for each y-array. The next two rows are self-explanatory.

Let $\bar{x}$ be the deviation of the average of an x-array from th average of the entire sample. From the following diagram, is seen that

$$\bar{x}'_{ax} = \bar{x} + \bar{x}'.$$



$X' = OX'$ = arbitrary value, $a_X = OM$ = average of all X' OA = average of any x-array. The deviation of a_X from X' $OM - OX' = \bar{x}'$, previously computed. $OA - OM = \bar{x}$. OA - $OX' = \bar{x}'_{ax}$. And $\bar{x}'_{ax} = \bar{x} + \bar{x}'$.

Then, in any one row of n items,

$$n\bar{x}'^2_{ax} = n(\bar{x} + \bar{x}')^2$$
$$= n\bar{x}^2 + 2n\bar{x}\bar{x}' + n\bar{x}'^2.$$

Then, for the entire table,

$$\Sigma n\bar{x}'^2_{ax} = \Sigma n\bar{x}^2 + 2\bar{x}'\Sigma n\bar{x} + \bar{x}'^2\Sigma n$$
$$= \Sigma n\bar{x}^2 + N\bar{x}'^2.$$

Since $\Sigma n\bar{x}$ equals summation for all rows of the total deviation from a_X of all items in each row, and so is the sum of the deviations from a_X for all N items, $\Sigma n\bar{x} = 0$.

From the above equation,

$$\frac{\Sigma n\bar{x}^2}{N} = \frac{\Sigma n\bar{x}'^2_{ax} - N\bar{x}'^2}{N}.$$

$\Sigma n\bar{x}^2$ is the sum of the weighted square deviations of the averages of the rows from a_X. Therefore, by definition,

$$\sigma^2_{mx} = \frac{\Sigma n\bar{x}^2}{N} = \frac{\Sigma n\bar{x}'^2_{ax}}{N} - \bar{x}'^2.$$

Similarly,

$$\sigma^2_{my} = \frac{\Sigma n\bar{y}^2}{N} = \frac{\Sigma n\bar{y}'^2_{ay}}{N} - \bar{y}'^2.$$

Now from the last column of Table LXX and the previous computed $\bar{x}'$,

$$\sigma^2_{mx} = \frac{55.22}{60} - (0.35)^2$$
$$= 0.9203 - 0.1225$$
$$= 0.7978.$$
$$\sigma_{mx} = \sqrt{0.7978} = 0.89.$$

Then,

$$\eta_{XY} = \frac{\sigma_{mx}}{\sigma_x} = \frac{0.89}{1.30} = 0.68.$$

In the same manner,

$$\sigma_{m_Y} = \sqrt{\frac{60.22}{60} - (-0.15)^2}$$
$$= 0.98.$$

And

$$\eta_{YX} = \frac{0.98}{1.51} = 0.65.$$

The f of the table is the same as the n of the formulas.

The quantity η_{XY} is the correlation ratio of X on Y, η_{YX} the correlation ratio of Y on X.

26. Degree of Departure from Linear Regression.—In every case η is greater than the numerical value of r unless there is strict linear regression giving $r = \pm 1$, when $\eta = r$. The difference $\eta^2 - r^2$ gives a measure of the degree of departure from linear regression. An approximate value of the probable error of $\eta^2 - r^2$ is given[1] as

$$2(.6745)\sqrt{\frac{\eta^2 - r^2}{N}} = 1.3490\sqrt{\frac{\eta^2 - r^2}{N}}.$$

If $\eta^2 - r^2$ is more than 2.5 times its probable error, it quite likely indicates a significant departure from linear regression. In other words, if

$$\frac{\sqrt{N}}{1.3490}\sqrt{\eta^2 - r^2} < 2.5,$$

regression may be regarded as linear.

Applying this test to

$$\eta_{XY} = 0.68,$$

$$\frac{\sqrt{60}}{1.3490}\sqrt{(0.68)^2 - (0.63)^2} = \frac{7.75}{1.3490}\sqrt{0.4624 - 0.3969} =$$
$$\frac{7.75\sqrt{0.0655}}{1.3490} = \frac{7.75 \times 0.26}{1.3490} = 1.5 < 2.5.$$

For $\eta_{YX} = 0.65$, the result, $0.92 < 2.5$, is obtained. This means that with the given correlation table it is permissible to regard regression as linear.

In the case of strict linear regression with RR' coinciding with SS',

$$\eta^2_{XY} = \eta^2_{YX} = r^2 = 1.$$

Even with strict linear regression, fluctuations due to sampling will likely make η^2 differ slightly from r^2.

[1] TRUMAN L. KELLEY, "Statistical Method," p. 239.

27. Correlation when $r = 0$.—In case the curve of regression is a periodic wave curve, r may be very small or zero, though there is perfect correlation. The correlation ratio is, then, much better to use for a measure of correlation. The correlation coefficient r, being smaller than the correlation ratio η, is more conservative than η. Sometimes it is too much so.

The coefficient r, together with σ_X and σ_Y, defines the slopes of the regression lines and so furnishes a means of determining general tendencies when regression is sufficiently close to linearity. For instance, the general rate of increase of leaf breadth with increase of length, or the general rate of increase of leaf length with increase of breadth in the universe represented by the given sample, is indicated.

28. Regression Equations Referred to Axes through $(0, 0)$.—The regression equation of x on y,

$$\bar{x} = r\frac{\sigma_X}{\sigma_Y}y,$$

gives the proportional change in the X variable to be expected with unit change in the Y variable. This equation is referred to

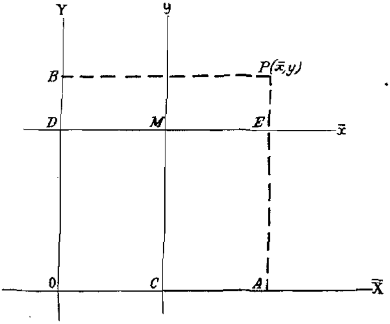


Fig. 55.

$\bar{x}$- and y-axes through the arithmetic means, a_X and a_Y, of the two variables. It is a simple matter to transform the equation to $\bar{X}$ and axes Y through the zeros of the two scales of original measurements of X and Y. In Fig. 55, the point P has coordinates $\bar{x}$ and y equal to ME and EP, referred to $\bar{x}$-axis and y-axis, and coordinates $\bar{X}$ and Y equal to OA and AP, referred to $\bar{X}$-axis

and Y-axis. Since O is at the zero of both scales and M at the arithmetic means, $OC = a_X$ and $OD = a_Y$. It is seen, at once, that

$$X = OC + ME = \bar{x} + a_X,$$

and

$$Y = OD + EP = y + a_Y.$$

Whence,

$$\bar{x} = X - a_X,$$

and

$$y = Y - a_Y.$$

To get the transformed equation, substitute these values of $\bar{x}$ and y in the original equation.

$$\bar{x} = r\frac{\sigma_X}{\sigma_Y}y.$$

becomes

$$\bar{X} - a_X = r\frac{\sigma_X}{\sigma_Y}(Y - a_Y)$$

or

$$\bar{X} = r\frac{\sigma_X}{\sigma_Y}Y + \left(a_X - r\frac{\sigma_X}{\sigma_Y}\cdot a_Y\right).$$

29. Example.—Applying this to the leaf lengths and breadths of Table LXX for which

$$\begin{aligned} a_X &= 26.04, \\ a_Y &= 36.25, \\ \sigma_X &= 3.90, \\ \sigma_Y &= 7.55, \\ r &= 0.63, \end{aligned}$$

the result is

$$\bar{X} = \frac{0.63 \times 3.90}{7.55}Y + \left(26.04 - \frac{0.63 \times 3.90}{7.55} \times 36.25\right).$$

or

$$\bar{X} = 0.33Y + 14.26.$$

This is the equation from which to estimate the average leaf breadth $\bar{X}$ to be expected for any given length Y on the assumption of linear regression.

For example, if length Y is 27, the expected average breadth $\bar{X}$ associated with it is

$$\begin{aligned} \bar{X} &= 0.33 \times 27 + 14.26 \\ &= 23.17. \end{aligned}$$

By inspection of the table, the average breadth of leaves in the sample of 60 associated with length 27 is 23.5.

Similarly, the regression equation

$$\bar{y} = r\frac{\sigma_Y}{\sigma_X}x$$

is transformed into

$$\bar{Y} = r\frac{\sigma_Y}{\sigma_X}X + \left(a_Y - r\frac{\sigma_Y}{\sigma_X}a_X\right).$$

For the table of leaf lengths and breadths, this becomes,

$$\bar{Y} = 1.22X + 4.49.$$

For breadth $X = 22$ the equation calls for an average length $\bar{Y}$ of

$$\bar{Y} = 1.22 \times 22 + 4.49$$
$$= 31.33.$$

Inspection of the table shows an average length of 31.5 for breadth of 22.

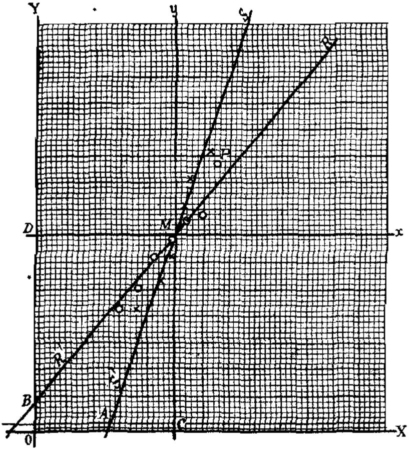


FIG. 56.—Regression lines of Y on X and X on Y.

30. Regression Lines Plotted.—These regression lines are plotted in Fig. 56. The line RR' shows the regression of y on x. It goes through the origin M for the equation

$$\bar{y} = 1.22x.$$

For the origin O, the equation is

$$Y = 1.22x + 4.49,$$

with the Y-intercept, $4.49 = OB$.

The line SS' shows the regression of x on y. It goes through the origin M for the equation

$$\bar{x} = 0.33y.$$

For the origin O the equation is

$$\bar{X} = 0.33Y + 14.26,$$

with the X-intercept,

$$14.26 = OA.$$

$$OC = a_X \text{ and } OD = a_Y.$$

31. Points Plotted from the Table.—The points marked with small crosses are determined from columns y' and $\bar{x}'_{ax}$ in Table LXX. The points marked with small circles are determined from rows x' and $\bar{y}'_{ay}$. In both cases, the corrections for $\bar{x}'$ and $\bar{y}'$ have been applied, and the results multiplied by the proper class interval. As an example for determining the points, take P, the uppermost circle. It is for $x' = 3$ and $\bar{y}'_{ay} = 2.50$ of the table.

$x = x' - \bar{x}' = 3.00 - 0.35 = 2.65$ class intervals
$x = 3 \times 2.65 = 7.95$ mm.
$\bar{y} = \bar{y}'_{ay} - \bar{y}' = 2.50 - (-0.15) = 2.65$ class intervals.
$\bar{y} = 5 \times 2.65 = 13.25$ mm.

The point $P \equiv (7.95, 13.25)$ is then plotted on the x-axis and y-axis. The other points are determined in the same way.

If either set of points be plotted, one may frequently be able to tell by experience whether the points group themselves sufficiently close to a straight line to warrant assuming linear regression without further computation.

Besides Pearson's coefficient r, and the correlation ratios η_{XY} and η_{YX}, other measures of correlation have been devised. One important coefficient of correlation is that from ranks.

32. Rank Correlation.—Rank is the position of an item in an array such that the item of rank 1 has no item preceding it. The item of rank 2 has one and only one preceding it. The item of rank 3 has two and only two items preceding it, and so on.

The formula for rank correlation is

$$\rho = 1 - \frac{6\Sigma d^2}{N(N^2 - 1)},$$

in which d stands for the difference in rank of the same item with respect to two different characteristics. N is the number of items or ranks.

If ranks with respect to two characteristics are regarded as the values, X and Y, of two associated variables, and if there are no duplications of rank, and no ranks lacking, this formula for ρ (rho) is the same as for Pearson's r. It thus assumes linear regression. It is frequently applied to other distributions with the risk of leading to serious inaccuracies. It should always be used with great caution.

33. Illustration.—As an illustrative example, take the following table of ranks of five students in English and in Mathematics, Table LXXI.

TABLE LXXI.—RANKS OF FIVE STUDENTS IN ENGLISH AND MATHEMATICS

Student	Rank		d	d^2
	English	Mathematics		
A	1	2	1	1
B	2	4	2	4
C	3	1	2	4
D	4	5	1	1
E	5	3	2	4
				$\Sigma d^2 = 14$

Mr. A ranked first in English and second in Mathematics; Mr. B ranked second in English and fourth in Mathematics; and so on.

$$\rho = 1 - \frac{6 \times 14}{5(25 - 1)} = 0.30.$$

This indicates that there was not very high correlation between English grades and Mathematics grades for these five students. This means that, with those students for which these five may be regarded as a representative sample, one cannot expect those getting high English grades necessarily to get high Mathematics grades, and vice versa.

34. Derivation of Formula.—Let X and Y be the ranks of the same item with respect to two characteristics; a_X and a_Y the corresponding respective arithmetic means; and σ_X and σ_Y their standard deviations. The X ranks, under the assumption of no

duplications, run through the series of natural numbers, 1, 2, 3, . . . N.

It is readily proved in algebra that the sum of the series

$$1 + 2 + 3 + \ . \ . \ . + N = \frac{N(N+1)}{2}$$

and

$$1^2 + 2^2 + 3^2 + \ . \ . \ . + N^2 = \tfrac{1}{6} N(N+1)(2N+1).$$

Then,

$$a_X = \frac{\Sigma X}{N} = \frac{1 + 2 + 3 + \ . \ . \ . + N}{N} = \frac{N+1}{2}.$$

Since Y has the same set of values,

$$a_Y = a_X = \frac{N+1}{2}.$$

The two arithmetic means are equal.

$$N\sigma^2_X = \Sigma(X - a_X)^2 = \Sigma X^2 - 2a_X \Sigma X + Na^2_X.$$

Substituting the values of ΣX^2 and ΣX from the above series, and the value of a_X

$$N\sigma^2_X = \frac{1}{6} N(N+1)(2N+1) - 2 \cdot \frac{N+1}{2} \cdot \frac{N(N+1)}{2} + \frac{N(N+1)^2}{4}$$

$$= \frac{N^3 - N}{12}.$$

Therefore,

$$\sigma^2_X = \frac{N^2 - 1}{12} = \sigma^2_Y.$$

Let $X - a_X = x$, deviation of X from arithmetic mean,
and

$Y - a_Y = y$, deviation of Y from arithmetic mean.

Subtracting,

$$X - Y = x - y,$$

Since

$$a_X = a_Y.$$

Now

$$\sigma^2_X = \frac{\Sigma x^2}{N},$$

$$\sigma^2_Y = \frac{\Sigma y^2}{N}.$$

Let

$$\sigma^2_{X-Y} = \frac{\Sigma(X - Y)^2}{N} = \frac{\Sigma(x - y)^2}{N}.$$

Now

$$\Sigma(x - y)^2 = \Sigma x^2 + \Sigma y^2 - 2\Sigma xy.$$

So

$$N\sigma^2_{X-Y} = N\sigma^2_X + N\sigma^2_Y - 2\Sigma xy. \qquad (1)$$

Pearson's coefficient is

$$r = \frac{\Sigma xy}{N\sigma_X\sigma_Y}.$$

So

$$\Sigma xy = Nr\sigma_X\sigma_Y.$$

Substituting in Eq. (1)

$$N\sigma^2_{X-Y} = N\sigma^2_X + N\sigma^2_Y - 2Nr\sigma_A\sigma_Y.$$

Solving for r,

$$r = \frac{\sigma^2_X + \sigma^2_Y - \sigma^2_{X-Y}}{2\sigma_X\sigma_Y}.$$

Since $\sigma_X = \sigma_Y$, substituting σ_X for σ_Y, giving

$$\begin{aligned} r &= \frac{2\sigma^2_X - \sigma^2_{X-Y}}{2\sigma^2_X} \\ &= 1 - \frac{\sigma^2_{X-Y}}{2\sigma^2_X} \\ &= 1 - \frac{\Sigma(X - Y)^2}{2N \cdot \frac{N^2 - 1}{12}} \\ &= 1 - \frac{6\Sigma(X - Y)^2}{N(N^2 - 1)} \\ &= 1 - \frac{6\Sigma d^2}{N(N^2 - 1)}, \end{aligned}$$

since $X - Y$ is d, the difference in rank. This expression for r is called ρ to distinguish it from r itself.

35. Duplicated Ranks.—If there are duplications of rank, what is known as the mid-rank method may be used. It is best shown by an illustration, using Table LXXII.

If the grades of C and D had not been alike but between 96 and 88, their ranks would have been 3 and 4, respectively. The mid-rank method assigns them each a rank halfway between 3 and 4, namely 3½. The grades of F, G, and H, being all 85, are treated the same way in ranking. If they had been different, the ranks would have been 6, 7, and 8. The mid-rank 7 is given to each. Then the formula for ρ is applied as before.

TABLE LXXII.—RANKS OF 10 STUDENTS IN ENGLISH AS SHOWN BY THEIR GRADES

Student	Grade	Rank
A	98	1
B	96	2
C	95	3½
D	95	3½
E	88	5
F	85	7
G	85	7
H	85	7
I	75	9
J	60	10

36. Error Due to Duplication.—In this way the arithmetic mean of ranks is preserved, but an error is introduced into the standard deviation. If there are few cases of duplication and only two duplicates in each case, the error is small. If there are many cases of duplication of rank or several individuals having the same rank, the error may lead to inaccurate conclusions.

See, for example, what error is introduced in standard deviation in the above example in Table LXXII. Standard deviation, computed from Table LXXIII is $\sqrt{8.00} = 2.828$.

TABLE LXXIII.—COMPUTATION OF STANDARD DEVIATION OF RANKS OF TABLE LXXII

Student	Rank	d_a	d^2_a
A	1	−4.5	20.25
B	2	−3.5	12.25
C	3.5	−2.0	4.00
D	3.5	−2.0	4.00
E	5	−0.5	0.25
F	7	1.5	2.25
G	7	1.5	2.25
H	7	1.5	2.25
I	9	3.5	12.25
J	10	4.5	20.25
Total..........	55	.....	80.00

$a = {}^{55}/_{10} = 5.5 \qquad \sigma^2 = {}^{80}/_{10} = 8.00$

Standard deviation by formula, when the ranks are all different, is

$$\sigma = \sqrt{\frac{N^2 - 1}{12}} = \sqrt{\frac{100 - 1}{12}} = \sqrt{8.25} = 2.872.$$

The error is about 2 per cent.

If only C and D had been of equal rank, σ^2 would have been 8.2, differing but little from $\sigma^2 = 8.25$, derived from the formula.

Of course, in case of decided error in σ, it is not very laborious to compute the correct value and it should be done.

In the formula for ρ, each value of the variable has been assigned to its proper rank. This may be expected to give a different value from the product-moment method where the actual grades would be used. For example, if two sets of values were as follows,

X	100	99	98	90	40	30	10	5
Y	99	98	97	96	95	94	93	90,

differences of rank are all zero, and hence $\rho = 1$.

It is at once seen that the standard deviation of X will be very much larger than that given by ranks. That for Y will be more nearly the same as that by ranks. The coefficient ρ will differ materially from r.

Comparison of ρ and r is hardly justifiable in many cases. One set of values for two associated variables may have the same differences in ranks as has another set of values, giving the same ρ in both cases, while the two values of r differ quite materially. In the above example, $\rho = 1$. Computing r, it is found to be 0.925. If the same values of X are used and the values of Y are each twice the corresponding values of X, then $\rho = 1$ and $r = 1$.

If it is not possible to put quantitative values on the variables, but they can be ranked, then resort to the use of ρ. Karl Pearson has derived the relationship

$$r = 2 \sin \left(\frac{\pi}{6} \rho\right)$$

when there is normal probability distribution in the actual scores. He comments that the method of obtaining r by means of ρ is less exact than the method developed in this chapter.

87. Graphical Determination of r.—Graphically, a rough value of r may be obtained by plotting the points for the regres-

sion lines, drawing the regression lines by the eye, and measuring their slopes b_1 and b_2. The product of these slopes gives r^2.

$$r = \sqrt{b_1 b_2}.$$

No further methods of measuring correlation will be taken up in this book.

38. Warning.—Too much stress cannot be put on using the coefficient of correlation, as well as other statistical constants, with care. Do not draw conclusions that are unwarranted. Remember that r might be small, though there is perfect correlation when regression is non-linear. On the other hand, r may be large and the variables have no actual connection. Table LXXIV gives an illustration.

TABLE LXXIV.—BANK CLEARINGS IN BILLIONS OF DOLLARS AND COMPOUND AMOUNT OF $1 AT 5 PER CENT

Date	Compound amount of $1 at 5 per cent, X	Bank clearings in billions, Y	x	y	x^2	y^2	xy
1883......	1.00	14	−1.40	−24.75	1.96	612.6	34.65
1893......	1.63	23	−0.77	−15.75	0.59	248.1	12.13
1903......	2.65	43	+0.25	+ 4.25	0.06	18.1	1.06
1913......	4.32	75	+1.92	+36.25	3.69	1,314.1	69.60
Totals....	9.60	155	......	....	6.30	2,192.9	117.44

$$a_X = \frac{9.60}{4} = 2.40. \qquad a_Y = \frac{155}{4} = 38.75.$$

$$r = \frac{117.44}{\sqrt{6.30 \times 2,192.9}} = 0.999.$$

This shows practically perfect correlation. Yet no one thinks that a dollar increases at compound interest because bank clearings increase. Nor can it be thought that bank clearings increase because of compound amount. In fact, a different set of dates for bank clearings would have given a very different result.

Correlation, as dealt with in this chapter, has picked out two variables and attempted to discover a probable relationship. As a matter of fact, the two variables are usually affected by a multitude of causes that have not been taken into account.

A high coefficient of correlation does not *prove* anything. It suggests the probability of a *cause-and-effect* relationship between the two variables. The statistician has no further concern. The

investigator, dealing with these variables, takes the suggestion and searches for a reason. Frequently, the statistician and the investigator are the same individual. He must be careful not to let his statistics run away with his reason.

Exercises

1. Compute Pearson's coefficient of correlation for each of Tables LXII to LXIX, inclusive. What is the reason for the increasing values of r?

2. Compute Pearson's coefficient of correlation for each of the Tables LXII to LXIX, inclusive, using the short-cut method as displayed in Table LXX.

3. Find η_{XY} and η_{YX} for each of the Tables LXII to LXIX inclusive.

4. Apply the test of Sec. 26 and determine whether there is sufficient departure from linear regression to be significant for any of the Tables LXII to LXIX.

5. Plot the regression lines for each of Tables LXVI to LXIX inclusive.

6. Using the frequency table of mean monthly temperatures at Seattle as found in Ex. 2, Chap. II, and the frequency table of monthly precipitation as formed in Ex. 3, Chap. II, construct a correlation table similar to Table LXX. Determine the arithmetic mean and standard deviation for mean monthly temperature and for monthly precipitation. Compute Pearson's coefficient of correlation and its probable error. What is the meaning of the negative sign of r?

7. Determine the regression coefficients of mean monthly temperature on monthly precipitation and of monthly precipitation on mean monthly temperature, using the data of Ex. 6.

8. Given the following associated values of X and Y, compute Pearson's coefficient of correlation between X and Y and its probable error.

X	3	5	8	12	17	23	30
Y	1	2	6	24	40	50	60

9. From the following table compute the coefficient of correlation between savings-bank deposits and strikes and lockouts in the United States over the period 1916 to 1922 inclusive.

Savings-banks Deposits and Strikes and Lockouts in United States

Date	Savings-banks deposits, in billions of dollars	Strikes and lockouts, in thousands
1916	5.1	3.8
1917	5.4	4.4
1918	5.5	3.3
1919	5.9	3.6
1920	6.5	3.3
1921	6.0	2.3
1922	7.2	1.0

10. Compute the probable error of r obtained in Ex. 9. What is the significance of the result?

11. Find Pearson's coefficient of correlation between leaf length and leaf breadth as shown in Table XXII. Use the classes of Tables XXIII and XXIV. Use the short-cut method. Compute the probable error of the coefficient.

12. Find Pearson's coefficient of correlation between leaf length and leaf breadth from the data of Ex. 18, Chap. II. Use the classes of the tables in Exs. 19 and 20, Chap. II. Compute the probable error of the coefficient.

13. Compute Pearon's coefficient of correlation between leaf length and leaf breadth from the data of Ex. 18, Chap. II. Use the same class intervals as were used in Exs. 19 and 20, Chap. II, but change the class limits. Compute the probable error. How do the results compare with those of Ex. 12?

14. Plot the regression lines for the data used in Ex. 12.

15. Plot the regression lines for the data used in Ex. 13.

16. Letting leaf breadth be the X variable, and leaf length be the Y variable, find η_{XY} and η_{YX} for the data of Ex. 12.

17. Similarly, find η_{XY} and η_{YX} for the data of Ex. 13.

18. Apply the test of Sec. 26 and determine whether there is sufficient departure from linear regression to be significant in the data used in Ex. 12.

19. Similarly, apply this test to the data of Ex. 13.

CHAPTER X

LOGARITHMIC GRAPHICAL REPRESENTATION

Non-uniform scales were briefly treated in Chap. III. It was ͻen that if two variables, x and y, were related by the equation

$$y = ka^x,$$

y plotting pairs of values of x and log y the resulting graph ould be a straight line. Look up and tabulate the logarithms of he values of y and plot these with the corresponding values of x. his is equivalent to making a uniform scale of logarithms for the -scale and plotting the values of log y on this scale. The best ay, however, is to have a non-uniform, logarithmic scale on hich to plot the values of y directly without having to look up garithms in a table.

Graph paper ruled to such a scale is now readily obtained on the arket.[1] Paper ruled to uniform scale one way and logarithmic ale the other way is called semi-logarithmic. It will be ssumed that such ruled paper is at hand.

1. Illustration.—With the following set of pairs of values of x nd y, the two methods may be illustrated:

x	10	20	30	40	50
y	1.791	3.207	5.743	10.286	18.420
log y	0.253	0.506	0.759	1.012	1.265

Figure 57 shows a uniform x-scale and a uniform log y-scale lotting the above pairs of values of x and log y, it may be seen at the points fall in a straight line. Now plotting the values of and y directly on semi-logarithmic paper, line AB (Fig. 58) is btained. This also proves to be a straight line. In either ig. 57 or 58, the fact that the line is straight shows uniform hange in log y. That is, for equal changes in x, the changes in

[1] Codex Book Company, New York, and the Standard Graphic Chart ompany, New York.

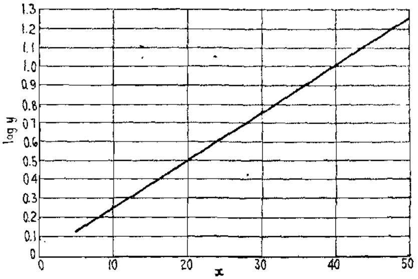


FIG. 57.—Uniform log y-scale.

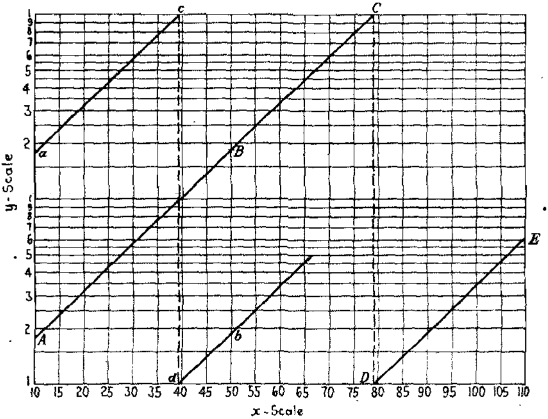


FIG. 58.—Compound amount. Logarithmic y-scale. Ratio chart.

$\log y$ are also equal. Or, the changes in $\log y$ are proportional to changes in x. Now

$$\log y_2 - \log y_1 = \log \frac{y_2}{y_1},$$

$$\log y_3 - \log y_2 = \log \frac{y_3}{y_2},$$

and so on.

If these successive differences of logarithms are equal, the ratios of the corresponding values of y are equal. If

$$\log y_2 - \log y_1 = \log y_3 - \log y_2,$$

then

$$\frac{y_2}{y_1} = \frac{y_3}{y_2}.$$

Since these figures give straight lines, it is seen at once that with equal changes of x (say 10 at a time) the changes in the corresponding ordinates to the line are equal. This means that the ratio of the corresponding values of y are equal. That is, the *rate of change in y is constant.* These facts might have been seen by inspection of the table of values. The variable x changes 10 at a time. The corresponding values of $\log y$ change 0.253 at a time. The ratio of each value of y to the preceding value is 1.791. For $x = 10$, $y = 1.791$. For $x = 20$, y is $1.791 \times 1.791 = 3.207$. For $x = 30$, y is $1.791 \times 3.207 = 5.743$, and so on. The values of x form an arithmetic progression, while the corresponding values of y form a geometric progression. For this reason, the graph of Fig. 58 is frequently called a *geometric chart,* those to uniform scale being called *arithmetic charts.*

2. Reading of Ratios.—Since $\log y_2/y_1 = \log y_2 - \log y_1$, the ratio of y_2 and y_1 may be determined, at once, graphically from the chart. With a divider, take off the difference of the ordinates at x_2 and x_1. Set this on the logarithmic scale with one leg of the divider at 1, and read 1.79 at the point on the scale where the other leg of the divider falls. The reading on the logarithmic scale gives the value directly, since the points on the scale are marked with the numbers of which the distances are logarithms. So y is multiplied by 1.79 for every increase of 10 in x. By taking the difference of ordinates for a difference of 1 in x, the rate of change in y per unit change in x will be determined. In this figure, the scale is too small to do this very accurately, but if care is used in using the divider to transfer the difference in ordinates

to the y-scale, it will be found that the result is about 1.06. This means that for an increase of 1 in x, y is multiplied by 1.06. In other words, the graph shows the compound amount of $1 at 6 per cent per annum compounded annually. The straight line AB may be extended and the compound amount read off for any number of years.

3. Scale Numbers by Decks.—Notice that the printed scale at the left-hand margin of the paper runs up from 1 to 9 and then repeats. The distance from 1 to 1 is the logarithm of 10 and equals the unit of the scale to which logarithms were plotted. A complete ruling from 1 to 1 is called a deck. The publishers of such ruled paper provide sheets with one deck, two decks, or more. The series of printed figures on each deck is the same. The logarithm of 1 is 0. Any 1 of any deck may be used as the 0, or starting point. If the bottom 1 on the sheet is the starting point, or zero, the next one above it is read 10, since it is at unit distance from the zero, and $\log 10 = 1$. The next 1 above that is read 100, since it is at two units distance from the zero, and $\log 100 = 2$. If there are three decks, the next 1 would be read 1,000, since $\log 1{,}000 = 3$, and so on.

4. Reading of Scale Numbers.—The lowest 1 may be read 10, meaning that the zero or starting point is the width of one deck below this point, and off the paper. The next 1 above is read 100, the next 1,000, and so on. In fact, the lowest 1 may be read as any integral power of 10, such as 1,000, 100, 10, 1, 0.1, 0.01, and so on indefinitely. Then the next 1 above is the next power of ten, and so on. If the lowest 1 is 100,000, the next above it is 1,000,000, the next above that, 10,000,000, etc. Or if the top 1 is 0.01, the next below is 0.001, the next below that is 0.0001, and so on.

5. Logarithmic Graph Unlimited.—With these facts in mind, the statement that the straight line AB (Fig. 58) may be extended for any number of years is literally true. If, in this figure, the line AB is extended to the top of the sheet, the compound amount read where it reaches the top line, at C, is $100 if the lowest 1 is $1. This is at a little beyond $x = 79$. From the point C, drop vertically down to the bottom line of the ruling at D. Call the 1 of that line 100 and extend the graph with the same slope it had before till it hits the right-hand margin at E, where $x = 110$. Carry this point horizontally to the left-hand margin, mark the x-scale 110 in place of 10, and extend the graph with its proper

slope till it reaches the top line. Proceed as before till a sufficient range is covered. Of course, if the graph is divided into too many parts, it becomes confusing.

From this graph may be read not only the compound amount at any date, but the date at which the amount has reached any given value. For example, in how many years will $1, compounded annually at 6 per cent per annum, amount to $10? At the point on the graph where y is 10, x is about 39.5. In about 39.5 years the compound amount of $1 at 6 per cent is $10.

6. Further Illustration.—Now tabulate and plot the compound amounts of $10 at 6 per cent.

x	10	20	30	40	50
y	17.91	32.07	57.43	102.86	184.20

Since the rate is the same, the ratio of amounts at any two given values of x will be the same for $10 as for $1. So the graph will be parallel to AB. It starts at a, proceeds to c, drops down to d, and then up to b, with the same slope as the first graph.

Since the graphs are parallel, the vertical distance between them is everywhere the same. Taking off, with the dividers, the vertical distance between the two graphs and reading it in the y-scale, it is found to read 10. Therefore, the second amount, at any date, is 10 times the first amount.

7. Contrast with Uniform Scale.—In order to contrast the use of the logarithmic scale with that of a uniform scale, the two sets of data are plotted on uniform scales in Fig. 59. The line AB goes up rather slowly. The line ab goes up much more rapidly with increasing rapidity as x increases. Almost anyone looking at this chart would say at once that the variable represented by ab is increasing more rapidly than that shown by AB. The logarithmic graph shows at a glance that they are increasing at the same rate.

What Fig. 59 does show at a glance is that the absolute increases for each 10-year period are increasing, and that for the upper curve, they are much greater than for the lower. The uniform scale chart, showing absolute differences, is sometimes called a *difference chart* in contrast to *ratio chart* for that with a logarithmic scale. If a comparison in *rate* of growth is desired, the difference chart is almost useless. While differences are not at once seen on the ratio chart, they are easily obtained by following over from points on the graph to the scale marks and reading

them. The business man, the biologist, the sociologist, and the economist are almost always concerned with comparisons in growth and rate of growth. Hence it is, that the ratio chart is being used more and more. When its simplicity and power are understood, it appeals very strongly to the investigator in almost any branch where graphical studies can be used.

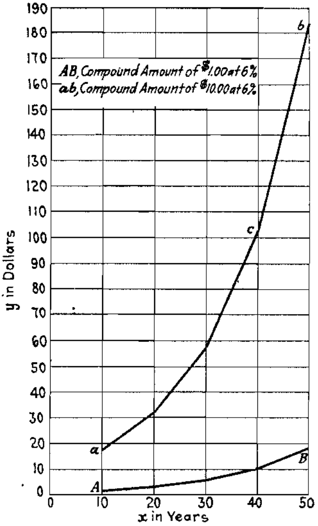


FIG. 59.—Compound amount. Uniform scales. Difference chart.

8. Organic Growth.—The equation

$$y = ka^{cx}$$

expresses what is called the law of organic growth. For equal differences in x, there are equal ratios in y. If x varies in arithmetic progression, y varies in geometric progression. This is shown by the simple case

$$y = 2^x$$

x	1	2	3	4	5	6
y	2	4	8	16	32	64

The x-series is an arithmetic progression with constant difference of 1. The y-series is a geometric progression with constant ratio of 2. Applied to a historical variable, the law states that at equal intervals of time the variable changes by a constant per cent. In the above example, if x is years, y doubles, increases 100 per cent each year. Many of the forces of nature follow more or less closely some form of this law. It is the law of compound amount. Business affairs have a tendency to follow this law. If salesmanship and internal management remained in the same state of efficiency, and external influences were constant, business might be expected to follow the law of organic growth rather closely.

9. Saturation Point.—As a matter of fact, it cannot be expected that any given business will follow this law indefinitely. There comes a slowing-up process, and what is called a *saturation point* is likely to be reached. The slowing-up process and the approach to saturation point are readily shown in a ratio chart. Take as an illustrative example the data of Table LXXV.

TABLE LXXV.—SALES IN A CERTAIN BUSINESS FOR 10 YEARS

Year	Sales	Increase	Per cent increase
1	$ 100,000		
2	200,000	$100,000	100
3	340,000	140,000	70
4	527,000	187,000	55
5	724,600	197,600	37.5
6	942,000	217,400	30
7	1,177,400	235,400	25
8	1,424,700	247,300	21
9	1,681,200	256,500	18
10	1,938,200	257,000	15

10. Illustration.—Plotted to uniform scale, these data give Fig. 60. It is evident that sales are increasing throughout the period. Since the curve gets steeper and steeper, it shows that the annual increase is getting larger from year to year throughout the period. This is also seen by inspection of the third column of the table. From the graph and the first three columns of the table, the inference might easily be drawn that this business is having a remarkable growth.

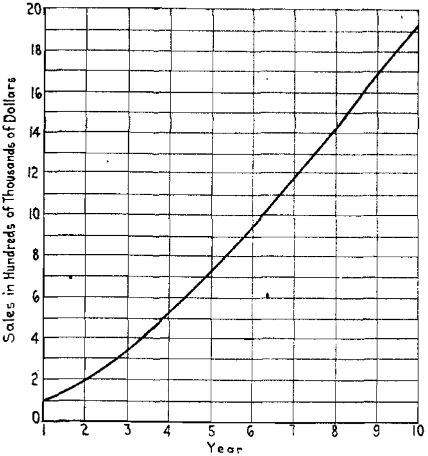


FIG. 60.—Sales for ten years. Uniform scales. Difference chart.

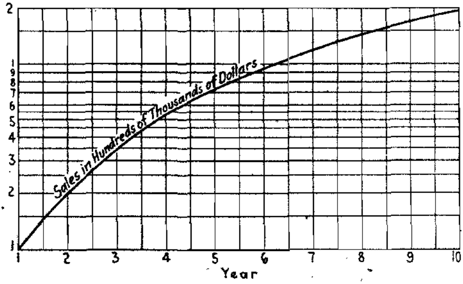


FIG. 61.—Sales for ten years. Ratio chart.

Figure 61 is a logarithmic graph of the same data. The curve goes up, showing an increase of sales every year. But the curve is concave downward; that is, it gets less and less steep with increase of time. This means that the per cent increase is getting less and less throughout the period. If that tendency continues, it may be expected that there will come a time when the curve will flatten out and cease to rise. When that time comes, the business will have reached a saturation point and will cease to grow unless efforts are made from within, or external conditions become more favorable. At any rate, it will be a critical situation, requiring careful treatment in order that the business may not decrease.

11. Forecasting.—Internal and external conditions remaining the same, a forecast may be made as to when the saturation point will be reached by drawing a smooth curve through the points plotted, extending it to the point where a tangent will be horizontal, and reading the corresponding date. It must be remembered in making such a forecast that it merely says that past tendencies indicate so and so. Change in market conditions, domestic or foreign demand, or quality of salesmanship may change the result completely. Controllable elements entering into the situation may be manipulated to keep the curve on its upward way. Or slackness in methods may turn the curve downward at an earlier date.

TABLE LXXVI.—MOTOR-CAR REGISTRATION IN THE UNITED STATES[1]

Year	Number of registrations
1912	1,010,399
1913	1,248,056
1914	1,768,963
1915	2,445,664
1916	3,512,996
1917	4,983,340
1918	6,146,617
1919	7,565,446
1920	9,231,941
1921	10,465,995
1922	12,238,375

[1] From "Statistical Abstract of the United States," except 1912, 1913, and 1914, which are from newspaper reports.

12. Example from Automobile Industry.—Figure 62 plotted from Table LXXVI, shows this tendency to approach saturation in the automobile industry. In 1913 and 1914 the curve grows steeper, showing there was a tendency to increase at an increasing rate. From 1914 to 1917 the curve is nearly straight, showing a tendency to maintain a constant rate of growth. From 1917 to 1920 the curve is again nearly straight, but less steep, showing a tendency to a constant rate of growth through the period but at a lower rate than from 1914 to 1917. In 1921 the rate of growth diminished still more. In 1922, however, the curve seems to be starting up a little more steeply.

A glance at the curve as a whole shows an undoubted tendency to approach a high point, or point of saturation. The slight

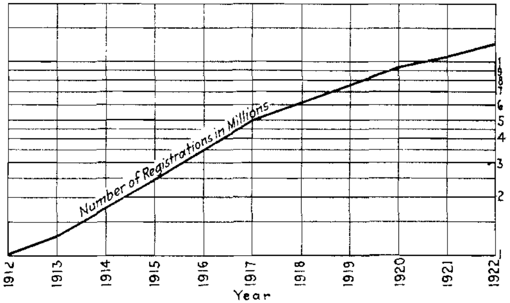


FIG. 62.—Motor-car registration in the United States. Ratio chart.

increase in rate of growth for 1922 may quite likely be temporary. Increased efforts at expansion may keep up the growth temporarily or even increase it. New elements of prosperity or increase of population may create sufficient growth in demand to keep up the rate of growth of business for a time. But it is impossible to get away from the general tendency of the 10-year period shown by the curve as a whole. This much is certain: The automobile industry is facing a probable saturation point in the not very distant future, and should conduct its business with that possibility in view. Later figures, only, will show the result.

13. Per Cent of Change.—Per cent decrease as well as per cent increase may be read directly from a ratio chart. Figure 63 is

plotted from data of Table LXXVII, showing the total production of gasoline and the total export in millions of gallons.

TABLE LXXVII.—PRODUCTION OF GASOLINE AND EXPORTS OF GASOLINE IN THE UNITED STATES[1]

Date	Production, in millions of gallons	Exports, in millions of gallons
1921:		
First quarter	1,268	155
Second quarter	1,305	136
Third quarter	1,268	112
Fourth quarter	1,312	130
1922:		
First quarter	1,315	141
Second quarter	1,513	168
Third quarter	1,656	140
Fourth quarter	1,718	131
1923:		
First quarter	1,823	194

[1] "Commerce Yearbook," 1922, U. S. Department of Agriculture.

Suppose it is asked what per cent the exports for the fourth quarter of 1922 were of the exports for the second quarter of that year. This would be obtained from the table by dividing 131 by 168. The ordinate at A is log 168, that at C is log 131; the log $^{131}/_{168}$ is log 131 − log 168 = BC. BC runs down and so is negative. Take off BC with dividers and transfer it to the 1 at the middle of the left-hand margin as the zero point, setting it off from 1 downward, since BC is negative. The reading obtained is 0.78. Therefore,

$$^{131}/_{168} = 0.78,$$

and the exports the fourth quarter were 78 per cent of those the second quarter.

To find what fraction (per cent) one quantity is of a larger one, take the difference of the ordinates of the two quantities, on a logarithmic scale, and measure it down from the top of a deck. The resulting reading gives the answer.

The exports for the first quarter of 1923 were how many times those of the fourth quarter of 1922? The table gives $^{194}/_{131}$.

$$\log {}^{194}/_{131} = \log 194 - \log 131 = CD$$

f the graph. *CD* runs up and is positive. With dividers, set off upward from 1 of the scale, getting a reading of 1.48. So 94 is 1.48 times 131, or is 148 per cent of 131.

To find how many times greater one quantity is than a smaller ne, take the difference of the ordinates of the two quantities, n a logarithmic scale, and measure it up from the bottom of a eck as zero point. The resulting reading gives the answer.

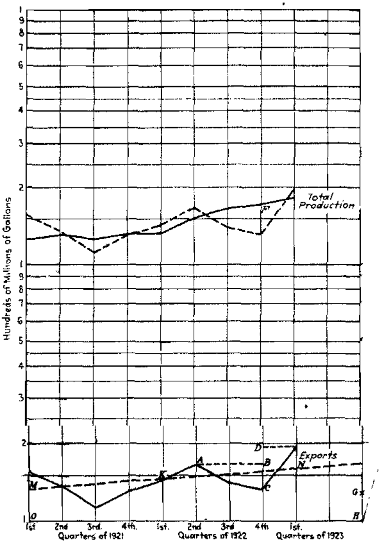


FIG. 63.—Production and exports of gasoline in the United States. Ratio chart. Least squares line of trend.

The exports for the fourth quarter of 1922 are what per cent the total production for that quarter? *FC* is the difference of dinates and is negative. Since *FC* is longer than the width of deck, it must be laid off from the top 1 of the scale. The result ads 76. This is not 0.76, but 0.076, since the logarithm *FC* s between −1 and −2, making −2 the characteristic of the garithm. So the answer is 7.6 per cent.

The total production for the fourth quarter of 1922 was how many times the exports for that date? Now, not *FC* but *CF* is used. It is laid off upwards from 1. The reading is 131. The production was 13.1 times the exports. As *CF* is greater than 1 and less than 2, the characteristic of the logarithm is 1.

If one is unfamiliar with the meaning of characteristic of a logarithm, the position of the decimal point may always be obtained by inspection of the numbers used. The problem of position of the decimal point is the same as it is on the slide rule.

Comparison of rates of change in these two variables is made more easy by moving the exports curve up one whole deck. The dotted line shows the resulting position of the graph. From the first to the second quarter of 1922, it is seen at once that the rate of change was nearly the same for both variables, because the two graphs are nearly parallel at this period. Throughout the entire period of time, it is easy to compare rates of change and see at what dates one rate was increasing and the other decreasing, or both changing in the same direction. It can also be seen which is changing at the greater rate by noting which graph is the steeper. If actual values of the two variables are to be considered, it must be remembered that the exports graph has been shoved up one deck so that the one for its scale is now the 10 for the production scale. If values of exports are read on the dotted graph, using the original scale, these values must each be divided by 10; that is, the decimal point must be moved one place to the left.

The correct readings are easily seen by proper labeling of the graphs. The total production graph is in hundreds of millions and the transposed exports graph is in tens of millions. The two graphs should be correspondingly labeled either along the graph or at one end.

14. Trend by Least Squares Line.—It may be desired to know the general tendency in rate of change of a variable. To find this, a least squares line may be fitted to the plotted points of the logarithmic graph.

Let

$$z = \log y.$$

Then in the least squares adjustment will be used x, z, x^2, and xz. Take the z-axis at the average $x = a_x$. Then, referring to the formulas derived in Chap. V

$$m = \frac{\Sigma x'z}{\Sigma x'^2}$$

and intercept

$$b = \frac{\Sigma z}{N} = a_z.$$

The data for computation of such a least squares line for gasoline exports are shown in Table LXXVIII.

x = number of time periods from the beginning of the record.
y = exports of gasoline in millions of gallons.
z = log y.
a_x = average of x.
$x' = x - a_x$ deviation of x from average x.

TABLE LXXVIII.—LOGARITHMIC LEAST SQUARES LINE FOR GASOLINE EXPORTS

x	y	$x' = x - a_x$	$z = \log y$	x'^2	$x'z$
0	155	−4	2.1903	16	−8.7612
1	136	−3	2.1335	9	−6.4005
2	112	−2	2.0492	4	−4.0984
3	130	−1	2.1139	1	−2.1139
4	141	0	2.1492	0	0.0000
5	168	1	2.2253	1	2.2253
6	140	2	2.1461	4	4.2922
7	131	3	2.1173	9	6.3519
8	194	4	2.2878	16	9.1512
36	. . .	. . .	19.4126	60	0.6466

$$a_x = {}^{36}\!/\!_9 = 4.$$

Note the positive and the negative quantities in summing $x'z$.

$$m = \frac{0.6466}{60} = 0.01078 = \log 1.025.$$

$$10m = 0.1078 = \log 1.282. \checkmark$$

$$b = \frac{19.4126}{9} = 2.1569 = \log 143.5.$$

The z-intercept b is the arithmetic mean of the logarithms of the values of y. Its antilogarithm, 143.5, is then the geometric mean of the values of the exports y. For average of rates, geometric mean should be used. The average x, $a_x = 4$, is the fourth date after the beginning of the period; that is, it is at first quarter

of 1922. It is simply the middate of the entire period. The value of b is laid off as an ordinate at this date. This is done by taking off with a divider the value of 143.5 on the logarithmic scale, setting it off as an ordinate at $x = 4$, first quarter of 1922, and locating the point K at the z-intercept. The actual distance is, of course, log 143.5, 2.1569. For greater accuracy in plotting, $10m$ is used to plot the slope. For a horizontal distance of 10 units of time, a rise is made of $10m = 1.28$ measured on the logarithmic scale.

$OH = 10$ units of time.
$HG = 1.28$ on the logarithmic scale.

From O to G gives the direction of the line. Draw a line through K, the z-intercept, parallel to a line through O and G. This line has the required slope and z-intercept. It is the least squares line sought. It is shown in the figure as the broken line MN. It may be called the "growth axis." The slope of the straight line connecting any two points in a logarithmic graph determines *average rate of change* between the two points over the period determined by them.

15. Meaning of Slope.—The slope $m = 1.025$ means that, on the average, exports of gasoline increased, or grew, 2.5 per cent per annum through the period tabulated. If the given data were graduated to this line, the exports for the first quarter of 1921 would have been about 130,000,000 gal., as shown by the ordinate to the point M at the beginning of the line of closest fit. At the end of the period, first quarter of 1923, the exports would have been about 159,000,000 gal. This is read directly at the point N. It may be calculated by

$$130(1.025)^8 = 158.5.$$

16. Forecasting.—If the actual points of the graph seem to follow this least squares line, line of closest fit, line of trend, fairly well, fluctuating from one side to the other, it may be used to forecast the approximate number of gallons to be exported at later dates. What may be said is, that if the trend of exports for the period graphed continues, the exports for future dates will be shown, on the average, approximately by this line of trend. A safe prediction can be made only for so long a period as this trend continues. Judgment, experience, and knowledge of external and internal conditions must be used to determine for how long a time the forecast may be used. Since so many

phenomena tend to follow the law of organic growth rather than a law of constant amount of increase, the logarithmic least squares adjustment is frequently much preferable to a least squares line on an arithmetic chart.

If plotting of the original data to uniform scales reveals that the *plotted points tend to group themselves about a straight line,* fluctuating from one side to the other, then constant difference changes are indicated. The trend is then shown by a line of closest fit determined on the constant difference chart.

If the points are plotted on semi-logarithmic paper and they seem to group themselves about a straight line, fluctuating from one side to the other, then the tendency to follow the law of organic growth is revealed. The trend line is plotted as a logarithmic least squares line. The values of ordinates to this line may be plotted on a uniform scale if a curve of closest fit is desired. This is the curve

$$y = k \cdot 10^{nx}$$

of page 199, Chap. VIII, in which 10 is the base of the logarithms used.

Many economic data, when plotted on a logarithmic scale, tend to follow a curve which, for a time, may be nearly straight but gradually tends to flatten out, showing the data are approaching a saturation point. It has been found that such data may frequently be fitted by a curve whose equation is

$$y = ak^{c^x}.$$

This is known as the "Gomperz curve." Taking logarithms of both sides gives

$$\log y = \log a + c^x \log k,$$

or

$$\log \frac{y}{a} = c^x \log k.$$

Again taking logarithms of both sides gives

$$\log \log \frac{y}{a} = x \log c + \log \log k.$$

Put $z = \log \log y/a$, $m = \log c$, and $b = \log \log k$ and the result

$$z = mx + b,$$

is which is linear in z and x.

If, then, for a set of observed values of x and y, the values of the logarithm of the logarithm of y/a are tabulated and these values are plotted with the observed values of x, and it is found

that the points tend to group themselves about a straight line, fluctuating from one side to the other, a Gomperz curve may be assumed to fit the original data. For the purpose of finding whether the data follow a Gomperz curve, a value for a may be assumed, since it is a constant multiplier. The fact that a Gomperz curve fits the data reveals that there is a tendency toward a saturation point. If a log log scale be constructed, points may be plotted directly.

It has been attempted to fit distribution of incomes to the equation,[1]

$$y = ax^m.$$

This equation was briefly discussed in Chap. VIII.

Taking logarithms of both sides gives

$$\log y = m \log x + \log a.$$

Put $z = \log y$, $v = \log x$, and $b = \log a$, and the equation takes the form

$$z = mv + b,$$

which is linear in z and v. To plot points directly for data supposed to fit this curve, logarithmic scales both ways are needed. The equation is known as "Pareto's law."

Since as numbers grow larger, their logarithms differ less and less (see any table of logarithms), a frequency distribution skewed in the direction of large values may frequently be made symmetrical by plotting the size of item to a logarithmic scale.

Recently paper has been printed having a uniform scale one way and a probability scale the other way. On this paper, a normal probability frequency distribution plots as a straight line. There is also printed paper with a logarithmic scale one way and a probability scale the other way. On this paper, a certain type of skewed frequency distribution plots as a straight line.

There are many applications of logarithmic and other non-uniform scales. Users of graphical work should become familiar with the making and interpretation of graphs to non-uniform scales. There is considerable literature on the subject.

[1] See "Income in the United States," vol. II, National Bureau of Economic Research.

Exercises

1. *Plot the following data on uniform, or arithmetic, scales.*

Year........	0	1	2	3	4	5	6	7	8	9	10
Sales, in *thousands of* dollars.	100	170	160	190	300	260	320	410	440	420	500

(*a*) Compute and plot a least squares straight line of closest fit.

(*b*) Plot the given data on semi-logarithmic paper with uniform scale for dates and logarithmic scale for sales.

(*c*) Reading the values of ordinates to the least squares line in (*a*), plot them on the logarithmic scale in (*b*). What is the significance of the curvature of the line thus determined?

2. Repeat the construction of 1(*b*). Compute and plot a logarithmic least squares straight line of closest fit.

Which is better for showing the trend of rate of change of sales, the line in 1(*c*) or the logarithmic least squares line just plotted? Which is better for showing the trend in per cent of change?

3. Plot the following data on uniform or arithmetic scales.

Year............	1915	1916	1917	1918	1919	1920	1921	1922	1923	1924
Amount, in dollars	1	3	2	6	4	12	8	24	16	48

Plot the same data on semi-logarithmic paper with uniform scale for the dates and logarithmic scale for the amounts.

What is the significance of the regular rise and fall of the logarithmic graph? Why is the other graph not also regular?

NUMBER OF AUTOMOBILES AND NUMBER OF DEATHS FROM AUTOMOBILES IN THE UNITED STATES

Year	Number of automobiles in use, *in millions*	Number of deaths from automobiles
1915	1.75	5,928
1916	2.42	7,397
1917	3.54	9,184
1918	4.94	9,672
1919	5.95	9,827
1920	7.90	8,878
1921	8.88	9,903
1922	10.51	11,066
1923	10.96	

4. Plot on semi-logarithmic paper the number of automobiles and the number of deaths from automobiles.

From the graphs thus plotted, determine and plot the number of deaths per thousand automobiles for the period 1915 to 1922.

On the assumption that the percentage decrease in number of deaths per thousand automobiles for 1923 would be the same as for 1922, continue the curve to 1923. From this, determine graphically the entry in the table for number of deaths in 1923. Continue the graph of number of deaths to 1923 by a dotted line.

5. Determine graphically the per cent increase in number of automobiles for 1920 and the per cent decrease in number of deaths for the same year.

Determine graphically the total per cent increase in number of automobiles for the period up to 1923. Determine the per cent increase of number of deaths for the same period.

6. Find graphically the average per cent increase per year in number of automobiles and the average per cent increase per year in number of deaths.

7. In January a certain business house introduced a new department. The monthly sales in this department for one year are shown in the following table.

Sales in New Department

Months	Sales in Thousands
January	$120
February	200
March	190
April	220
May	300
June	290
July	320
August	400
September	390
October	430
November	500
December	490

Plot the data of the table on millimeter paper. Label this Fig. 1. Determine the best period for a moving average. Compute the moving average and record the results in an additional column in the table. Show the trend for the year by plotting the moving average in Fig. 1.

Now plot the sales data and moving average on semi-logarithmic paper. Label this graph Fig. 2. Do not smooth the graphs of either figure. Draw the moving average graphs in dotted lines.

Do the points on the moving average fall nearly on a straight line in Fig. 1? What does this trend indicate in regard to the growth of the new department?

The line of trend in Fig. 2 is a curve becoming less and less steep toward the right. What does this indicate in regard to the growth of the new department?

If the dotted curve of Fig. 2 continues to curve downward at about the same general rate, what will happen to the rate of increase in sales in about

8. Find the least squares straight line of closest fit to the logarithmic graph of Ex. 7, Fig. 2. Interpret the meaning of this line as compared with the trend line of Fig. 2.

OPERATING REVENUES AND TAXES OF CLASS I RAILROADS IN UNITED STATES

Year	Total operating revenues in billions	Taxes in hundreds of millions
Ending June 30:		
1913	$3.11	$1.18
1914	3.03	1.35
1915	2.87	1.33
1916	3.38	1.46
Ending December 31:		
1916	3.60	1.57
1917	4.01	2.14
1918	4.88	2.23
1919	5.14	2.33
1920	6.18	2.72
1921	5.52	2.76
1922	5.56	3.01

9. Plot the data of the above table of operating revenues and taxes on millimeter paper. Connect each point to the next by a straight line, not drawing a smooth curve. Select a scale that will include all the data. Call this Fig. 1. Label each graph showing what it represents.

Now plot the data on semi-logarithmic paper. Plot the actual figures of the table. Label each graph showing what it represents and the scale used. Call this Fig. 2.

10. By inspection of Fig. 1, Ex. 9, would you judge that the fluctuations in taxes and in total operating revenues are at all alike? Do they increase together?

By inspection of Fig. 2, Ex. 9, would you judge that the fluctuations in taxes and in total operating revenues are at all alike? Is your answer the same as to the first question? If not, what is the reason for a difference in judgment?

11. Determine from Fig. 2, Ex. 9, the per cent taxes are of total operating revenues for the year 1917. What is the easiest way of doing this?

For the year 1920 total operating revenues were how many times the taxes? How is this most readily determined from the graph?

12. As shown by the graph of Ex. 9, how long after June 15, 1915, did it take total operating revenues to double? How is the result obtained? Did taxes double in this same period? If not, what per cent did they increase?

13. At how long a time before Dec. 31, 1920, were taxes 50 per cent what they were at that date? How is this most easily obtained from tl graph of Ex. 9?

14. Does the graph of Ex. 9 show any time when taxes and total operatic revenues were changing at the same rate? What shows this? Does th show in Fig. 1?

15. As shown by the graph, the total operating revenues for 1920 we what per cent of those for 1919? Also what per cent of those for 192 How are these results obtained?

CHAPTER XI

INDEX NUMBERS

Index numbers, in their simplest form, are simply the ratios, expressed as per cents, of each of a set of values of a variable to the value of some one number used as a base. The base usually is one of the set of numbers used, or the average of all, or the average of a group selected from the entire set.

1. Example.—The use of such index numbers as above defined is to facilitate comparisons in the *relative* values of two or more

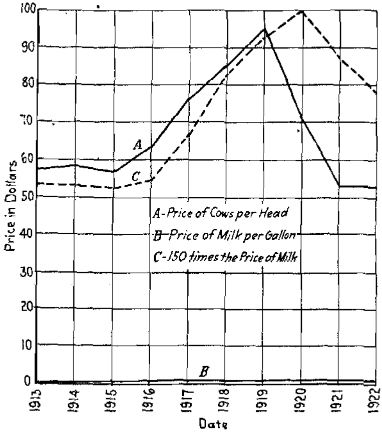


FIG. 64.—Prices of cows and milk.

sets of variables or of a single variable at different times. Table LXXIX gives the prices of cows and of milk over a series of years from 1913 to 1922. These prices are plotted in Fig. 64. Curve *A* shows the price of cows and *B* shows the price of milk, both in the same unit of measurement, dollars.

TABLE LXXIX.—PRICES OF COWS AND MILK[1]

Date	Cows per head on Dec. 15	Milk, in gallons, average	Index price of cows	Index price of milk	150 times price of milk
1913	$57.19	$0.356	85	74	53.4
1914	58.23	0.356	87	74	53.4
1915	56.79	0.352	85	73	52.8
1916	63.18	0.364	94	75	54.6
1917	76.16	0.442	114	92	66.3
1918	85.78	0.556	128	115	83.4
1919	95.54	0.620	143	128	93.0
1920	70.42	0.668	105	139	100.2
1921	53.30	0.584	80	121	87.6
1922	53.21	0.524	80	109	78.6
Average.......	66.98	0.482	100	100	

[1] "Statistical Abstract of United States" for 1922.

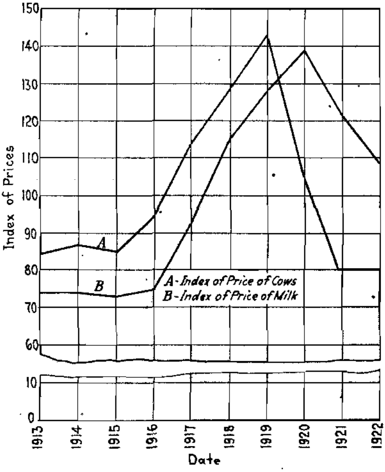


FIG. 65.—Indices of prices of cows and milk.

While curve *A* displays the changes in price of cows very well, curve *B* is so flat as to indicate almost no change at all in price of milk. If *B* had been plotted to a scale to show its variations, *A* would go off the paper. So, for the purposes of comparison of the variations in the two sets of prices, each set is reduced to percentage of the average price. These percentages are shown in the fourth and fifth columns of the table, and are called indices of the prices. These numbers are purely relative and completely independent of the units of measurement used in giving either set of prices. These indices are plotted in Fig. 65, curve *A* for index of price of cows and curve *B* for index of price of milk. It can now be seen at a glance how one set of relative prices changed as compared with the other.

The two sets of prices seem to go up and down together, except that the price of milk continued to rise in 1920 after the price of cows began to drop. From 1917 on, the price of milk seems to *lag* behind the price of cows. If that part of curve *B* were moved to the left one year, the two curves would nearly coincide for the remainder of the period. From 1920 on, the price of milk would seem to be unduly high as compared with the price of cows. In 1922 the price of cows did not drop. Perhaps the maintenance of relatively high price for milk stopped the decline in the price of cows. Of course, these data should be considered in connection with other data not recorded here.

2. Other Devices.—Sometimes such sets of numbers are compared by plotting one set to one scale and the other to another. This is a rather awkward way of accomplishing the result. The different scales have to be borne in mind.

Another method is to multiply one set by a number that will bring the two sets of values to comparable results. Thus the last column of Table LXXIX shows 150 times the price of milk. The number 150 was chosen because the average price of cows is about 150 times the average price of milk. This column of figures may be regarded as the price of 150 gallons of milk. The dotted curve *C* (Fig. 64) shows these new values. Curves *A* and *C* are easily compared.

3. Comparison of Average Values.—The purpose of index numbers, in general, is the same as that of the special form just discussed, that is, to reduce series of data, more or less complex, to numbers purely relative, which enables direct comparisons to be made. The general index number deals with averages and is a

ratio comparing the average value of a set of numbers at one time or place with the average value at some other time or place. This, on the face of it, seems like a simple matter. But when the various kinds of numbers to be compared, the different sorts of averages that may be used, the different methods of weighting possible, and the tests that index numbers should satisfy are considered, it is found that the problem of making a scientifically satisfactory index number is a rather complex one. It is a problem upon which the experts are not wholly agreed. Only the simpler aspects of the problem may be taken up in this work. To any one desiring a very complete survey and analysis of the subject, Prof. Irving Fisher's "The Making of Index Numbers" is recommended.

4. Definition.—Suppose an index of the prices of those commodities that enter cost of living is desired, showing the relative change in cost of living from one time to another. If all prices went up or down at the same rate and the quantities of commodities used did the same, the problem would be a simple one. But some prices change rapidly, some slowly. Some prices increase while others decrease. The quantities used vary in more or less irregular manner also. This complicates the problem.

The example just mentioned is an index of changes in costs from one time to another. Instead, the changes from one place to another might be shown. Instead of costs, there might be considered changes in values of goods produced, quantities of manufactured articles produced, prices of bonds, rates of exchange, wages, wholesale prices, retail prices, and many other things in regard to changes from one time to another or from one place to another. In fact, *the index number serves to make comparison of the sizes of quantities in one set of things under one situation with sizes of the quantities in the same set of things under another situation.* For the purposes of this chapter, the sizes of quantities compared will be prices of commodities in a set, and the different situations under which they are compared will be different times. The time unit taken will be the year. Prices of a set of commodities one year will be compared with prices of the same set another year. The year with which the comparison is made is called the *base year.*

5. Notation.—The following notation will be used.

p_o = price for base year.
p_k = price for the kth year.
q_o = quantity for base year.
q_k = quantity for the kth year.
p' = price of the first commodity in the set.
p'' = price of the second commodity in the set.
p''' = price of the third commodity in the set.
$p^{(r)}$ = price of the rth commodity in the set.
q' = quantity of the first commodity produced.
q'' = quantity of the second commodity produced.
q''' = quantity of the third commodity produced.
$q^{(r)}$ = quantity of the rth commodity produced.

Thus, p_3'' is the price of the second commodity in the set, for the third year, and q_4' is the quantity of the first commodity in the set produced in the fourth year; Σp_k = sum of the prices of the commodities for the kth year, and $\Sigma p_k q_o$ = sum of the products of price of each commodity for the kth year, multiplied by its quantity produced in the base year. That is,

$$\Sigma p_k q_o = p_k' q_o' + p_k'' q_o'' + p_k''' q_o''' + \ . \ . \ . \ + p_k^{(n)} q_o^{(n)},$$

if there are n commodities in the set.

Sometimes, instead of a base year, the average over the entire period, or the average for a decade or other period of time, is used as a base. In the case of place-to-place comparison, q_o and q_o would be price and quantity produced at the place taken as base with which comparisons are made. Or p_o and p_o may be taken as the average price and quantity for all the places considered.

The index number is nearly always expressed as a per cent. Sometimes it is expressed as a per thousand, as is done by G. H. Knibbs in Australia. *The London Economist* uses a per 2,200, there being 22 commodities in the set used.[1] This saves dividing summations by 22.

6. Average Price Relative.—The ratio of a price of a commodity at one time to its price at another time is called a *price relative.* Thus, p_k'/p_o' is the price relative of the first commodity for the kth year as compared with the base year. The average price

[1] Prof. Irving Fisher, "Making of Index Numbers," p. 371.

relative of n commodities is called an index number of prices. It equals

$$\frac{\sum \frac{p_k}{p_o}}{n}$$

7. Example.—The following two tables furnish data for illustrative problems.

TABLE LXXX.—PRODUCTION OF GRAINS IN THE UNITED STATES

In millions of bushels[1]

Date	Corn	Wheat	Oats	Rye	Barley	Buck-wheat	Rice (rough)
1913	2,447	763	1,122	41	178	14	26
1915	2,995	1,026	1,549	54	229	15	29
1917	3,065	637	1,593	63	212	16	35
1919	2,811	968	1,184	75	148	14	42
1921	3,069	815	1,078	62	155	14	38

[1] From "Statistical Abstract of the United States for 1922," pp. 166, 167.

TABLE LXXXI.—PRICE OF GRAINS IN THE UNITED STATES, IN CENTS PER BUSHEL

Farm values as of Dec. 1[1]

Date	Corn	Wheat	Oats	Rye	Barley	Buck-wheat	Rice (rough)
1913	69.1	79.9	39.2	63.4	53.7	75.5	85.8
1915	57.5	91.9	36.1	83.4	51.6	78.7	90.6
1917	127.9	200.8	66.6	166.0	113.7	160.0	189.6
1919	134.5	214.9	70.4	133.2	120.6	146.1	266.6
1921	42.3	92.6	30.2	69.7	41.9	81.2	95.2

[1] From "Statistical abstract of the United States for 1922," pp. 166, 167.

Using 1913 as base year, the price relatives of the grains tabulated in Table LXXXI are as follows. Each ratio is multiplied by 100.

TABLE LXXXII.—PRICE RELATIVES OF GRAINS IN THE UNITED STATES[1]

Grain	1913	1915	1917	1919	1921
Corn	100	83	185	195	61
Wheat	100	115	252	269	116
Oats	100	92	170	180	77
Rye	100	132	262	210	110
Barley	100	96	212	224	78
Buckwheat	100	104	212	193	107
Rice	100	105	221	311	111
Total	700	727	1,513	1,582	660

[1] Divisions by slide rule.

The price relative of corn for 1915 is $57.5/69.1 \times 100 = 83$. For rye for 1917, $166.0/63.4 \times 100 = 262$. The index numbers of prices of grains are:

1913 . . . . $\frac{700}{7} = 100$

1915 . . . . $\frac{727}{7} = 104$

1917 . . . . $\frac{1,513}{7} = 216$

1919 . . . . $\frac{1,582}{7} = 226$

1921 . . . . $\frac{660}{7} = 94$

8. Interpretation.—The interpretation is that for these grains, according to these index numbers, in 1915 the general average price was 4 per cent higher than in 1913; for 1917, 116 per cent higher than in 1913; for 1919, 126 per cent higher; and for 1921, 6 per cent lower than in 1913.

9. Weighted Averages.—This index number of prices is only one of many possible index numbers, some good, some bad. No account is here taken of the quantities produced. Each year there was much more corn produced than there was rye. A rise in the price of corn is more important than a rise in the price of rye. So, instead of a simple arithmetic mean, a weighted arithmetic mean suggests itself, weighting each price by the quantity produced. For the base year, there would be a weighted average of

$$\frac{\Sigma p_0 q_0}{\Sigma q_0} = w_0.$$

For year k,

$$\frac{\Sigma p_k q_k}{\Sigma q_k} = w_k.$$

The index for year k would be

$$\frac{w_k}{w_0} = \frac{\Sigma p_k q_k}{\Sigma q_k} \div \frac{\Sigma p_0 q_0}{\Sigma q_0}$$
$$= \frac{\Sigma p_k q_k}{\Sigma p_0 q_0} \div \frac{\Sigma q_k}{\Sigma q_0}.$$

This is Fisher's formula 52.

Another form that would suggest itself is obtained by weight each price relative by the value of the corresponding commod produced in the base year, that is, by $p_0 q_0$.

This formula would be

$$\frac{\Sigma p_0 q_0 \frac{p_k}{p_0}}{\Sigma p_0 q_0}.$$

This is Fisher's formula 3.

It reduces to

$$\frac{\Sigma p_k q_0}{\Sigma p_0 q_0},$$

which is Fisher's formula 53. In this form, it appears that quantities produced in the base year are used as weights for b years. This formula is used by the United States Bureau Labor.[1]

10. Other Averages Used.—The fact that there are other av ages than the arithmetic mean suggests their use in making in numbers. Professor Fisher has discussed 134 different in number formulas divided into six classes according to the t of average used. The only averages he uses are arithme harmonic, geometric, median, mode and aggregative.

The formula for the simple arithmetic index number already been given as

$$\frac{\Sigma \frac{p_k}{p_0}}{n}.$$

[1] PROF. IRVING FISHER, "The Making of Index Numbers."

The simple harmonic index number formula is

$$\frac{n}{\Sigma \frac{p_0}{p_k}},$$

being the harmonic mean of the price relatives.

The simple geometric index number is given by

$$\sqrt[n]{\frac{p_k'}{p_0'} \cdot \frac{p_k''}{p_0''} \cdot \frac{p_k'''}{p_0'''} \cdot \quad . \quad . \quad . \quad \cdot \frac{p_k^{(n)}}{p_0^{(n)}}} = \sqrt[n]{\Pi \frac{p_k}{p_0}}.$$

It is the geometric mean of the price relatives.

The simple median index number is the median of the price relatives.

The simple mode index number is the mode of price relatives, that one which occurs oftenest.

The formula for simple aggregative index number is

$$\frac{\Sigma p_k}{\Sigma p_0}.$$

It is the ratio of aggregate price of all the commodities for the kth year to their aggregate price for the base year, multiplied by 100.

The other 128 formulas are derived from these six by various systems of weighting, by "crossing" formulas, and by comparison with the value ratio

$$\frac{\Sigma p_k q_k}{\Sigma p_0 q_0}.$$

11. Systems of Weighting.—Professor Fisher uses four systems of weights:

I. $p_0 q_0$.
II. $p_0 q_k$.
III. $p_k q_0$.
IV. $p_k q_k$.

Other systems of weighting may easily be devised.

Using I, the simple arithmetic formula becomes

$$\frac{\Sigma p_0 q_0 \frac{p_k}{p_0}}{\Sigma p_0 q_0},$$

Fisher's 3. It has just been seen to reduce to Fisher's 53.

II gives

$$\frac{\Sigma p_0 q_k \frac{p_k}{p_0}}{\Sigma p_0 q_k},$$

Fisher's 5. It reduces to

$$\frac{\Sigma p_k q_k}{\Sigma p_0 q_k},$$

Fisher's 59 of the aggregate type.

III gives

$$\frac{\Sigma p_k q_0 \frac{p_k}{p_0}}{\Sigma p_k q_0}.$$

Fisher's 7.

IV gives

$$\frac{\Sigma p_k q_k \frac{p_k}{p_0}}{\Sigma p_k q_k},$$

Fisher's 9.

The application to other types of averages will not be recited here.

Among so many index numbers confusion is naturally expected. They do not give the same results. Some, then, must be better than others in giving a fair and just comparison of the fluctuations in average price, or whatever the numbers are applied to.

12. Three Tests.—There are certain tests that an index number should satisfy in order to be perfectly satisfactory. They are:

The commodity reversal test.
The time reversal test.
The factor reversal test.

13. The commodity reversal test is that the order of the commodities may be reversed, or, in fact, that they may be placed in any order whatever without making any difference in the result. This simply means that the superscripts on the letters

$$p', p'', p''', p^{iv}, \ldots p^{(n)}$$

may be rearranged in any manner desired. It is evident that this test is satisfied by all the formulas, since the sum of the price relatives, the product of the price relatives, the sum of their reciprocals, their median, their mode, their aggregate are none of them affected by change in the order of their occurrence.

14. The time reversal test interchanges the subscripts k and 0. That is, if, of two years, the ratio of the average price for the

second year to that of the first year is found, the interchange will give the ratio of the average price for the first year to that for the second year. The base year is simply interchanged with the other year, and the other year is made the base year. The resulting ratio (not index number) should be the reciprocal of the ratio obtained before the change was made. The product of the two should be unity. The product of the two resulting index numbers divided by 100 should be 100 per cent. This is evident, for, if with, say, 1900 as base, it is found that for 1905 the average price has doubled, the index for 1905 on 1900 as base is 200. Of course, if the average price has doubled from 1900 to 1905, then the 1900 average price ought to be half that for 1905 and the index for 1900 on 1905 as base would be 50. The product of the two ratios is

$$2/1 \times 1/2 = 1.$$

The product of the two index numbers divided by 100 is

$$\frac{200 \times 50}{100} = 100.$$

If there is any doubt in this argument in regard to going forward or backward in time, use index numbers of prices at one place as compared with prices at another place. Of two cities, A and B, there is no inherent reason for using one for the base rather than the other. Suppose the index of prices at A is obtained with the average price at B as base and found to be 200. That is, prices at A average twice as much as at B. Then prices at B must average half as much as at A. The index of prices at B, using average price at A as base, must be 50, if a satisfactory method of getting the index has been used.

15. Factor Reversal Test.—Just as the commodity reversal test interchanges the superscripts of the letters, and the time reversal test interchanges the subscripts, so the factor reversal test interchanges the two factors p and q. An index number satisfying the commodity reversal test remains unchanged when the test is applied. This must be expected, because Σp and Σq each remain unchanged, no matter what the order of summation. The index number satisfying the time reversal test becomes its own reciprocal, when the test is applied. This is to be expected, because

$$\frac{p_k}{p_0} \times \frac{p_0}{p_k} = 1$$

for each individual commodity. A fair index of average prices, then, should satisfy the same condition.

The ratio of total values,

$$\frac{\Sigma p_k q_k}{\Sigma p_0 q_0},$$

is a fixed determinate constant no matter what sort of index numbers may be used. Now,

$$\frac{p_k}{p_0} \times \frac{q_k}{q_0} = \frac{p_k q_k}{p_0 q_0}$$

for each individual commodity. So the ratio of *average* prices for two years multiplied by the ratio of *average* quantities produced for the same two years ought to give the ratio of total values. An index number of price, properly weighted, gives the general average ratio of prices for the year k as compared with the year 0. Also, an index number of quantity produced, properly weighted, gives the general average ratio of quantities for the year k as compared with the year 0. If, in a formula for an index number of price weighted according to quantity, the letters p and q are interchanged, the result is the index number of quantity weighted according to price. The factor reversal test requires that the product of the two resulting index numbers, when p and q are interchanged, should equal the ratio of total values.

Bear in mind that the formulas for index numbers and these three reversal tests apply not only to price and quantity but to any two sets of numbers the notation for which conforms to the notation used here. For example, p may be freight rates of a certain schedule of goods, p_k for water-borne commerce, and p_0 for rail, while q_k and q_0 are the numbers of tons shipped of the same goods by water and by rail respectively. Then $\Sigma p_k q_k$ is the total amount of freight paid for water-borne commerce and $\Sigma p_0 q_0$ for rail. Or, p may be the number of rows of corn on an ear and q the number of kernels in a row (assuming the same number in each row on any one ear). Let p_k and q_k be for Yellow Dent, and p_0, q_0 for Country Gentleman. Let p', q' be for the shortest ear in a random sample of n ears, p'', q'' for the next longer, p''', q''' for the next, and so on to $p^{(n)}$, $q^{(n)}$ for the longest of the set. In this case $\Sigma p_k q_k$ is the total amount of Yellow Dent corn and $\Sigma p_0 q_0$ is the total amount of Country Gentleman.

Manifestly, all three reversal tests ought to apply to the corn, in the above example. For the same reason, they should apply to price and quantity.

Now examine these tests as they apply to some of the formulas already given. The commodity reversal test may be omitted from the discussion for obvious reasons previously given.

16. Test of Average Price Relative.—Take the simple arithmetic mean of price relatives

$$\frac{\Sigma\frac{p_k}{p_0}}{n}.$$

This formula has been used a great deal by different sources of index numbers.

Applying the time reversal test, the result should be

$$\frac{\Sigma\frac{p_k}{p_0}}{n} \times \frac{\Sigma\frac{p_0}{p_k}}{n} = 1.$$

For simplicity take only two commodities. Then,

$$\frac{\frac{p_k'}{p_0'} + \frac{p_k''}{p_0''}}{2} \times \frac{\frac{p_0'}{p_k'} + \frac{p_0''}{p_k''}}{2} = \frac{\frac{p_k'}{p_0'} \times \frac{p_0'}{p_k'} + \frac{p_k''}{p_0''} \times \frac{p_0''}{p_k''} + \frac{p_k'}{p_0'} \times \frac{p_0''}{p_k''} + \frac{p_k''}{p_0''} \times \frac{p_0'}{p_k'}}{4}$$

is obtained.

The first two terms of the numerator are each 1. Neither of the other terms of the numerator equal 1 unless $p'' = cp'$ for each year. So unless the prices change proportionally, the product of the two index numbers cannot be expected to equal 1.

Take the following simple example of two commodities, A and B.

Prices			Quantities		
	1900	1905		1900	1905
A	20	40	A	3	6
B	100	50	B	4	2

The simple arithmetic index of prices of 1905 on those of 1900 as base is

$$\frac{\sum \frac{p_k}{p_0}}{n} = \frac{\frac{40}{20} + \frac{50}{100}}{2} = \frac{2 + \frac{1}{2}}{2} = 1\frac{1}{4} = 125 \text{ per cent.}$$

This makes the general average price of these commodities 25 per cent higher in 1905 than in 1900.

Applying the time reversal test and using 1905 as a base, the result would be

$$\frac{\frac{20}{40} + \frac{100}{50}}{2} = \frac{\frac{1}{2} + 2}{2} = 1\frac{1}{4} = 125 \text{ per cent.}$$

According to this, the general price of these commodities in 1900 was 25 per cent higher than in 1905. The absurd conclusion is thus made that each year the average price was 25 per cent higher than it was the other year. Of course, this was an extreme case in which half of the commodities doubled in price and the rest decreased one-half. The algebra of the formula has shown, however, that the two indices cannot be expected to be reciprocals of each other, unless prices change at a constant ratio. If this happened, there would be no need of index numbers anyway.

It follows that the simple arithmetic mean of price relatives is a very poor form of index number. The factor reversal test would give

$$\frac{\sum \frac{q_k}{q_0}}{n}.$$

$$\frac{\sum \frac{p_k}{p_0}}{n} \times \frac{\sum \frac{q_k}{q_0}}{n}$$

does not equal

$$\frac{\Sigma p_k q_k}{\Sigma p_0 q_0},$$

and the factor reversal test is not satisfied.

16. Tests of Other Formulas.—Consider Fisher's formula 52 (p. 280):

$$\frac{w_k}{w_0} = \frac{\Sigma p_k q_k}{\Sigma q_k} \div \frac{\Sigma p_0 q_0}{\Sigma q_0}.$$

For the time reversal, interchange the subscripts k and 0. This gives

$$\frac{\Sigma p_0 q_0}{\Sigma q_0} \div \frac{\Sigma p_k q_k}{\Sigma q_k} = \frac{w_0}{w_k}.$$

It is seen, at once, that the second result is the reciprocal of the first.

$$\frac{w_k}{w_0} \times \frac{w_0}{w_k} = 1.$$

The time reversal test is satisfied.

For the factor reversal test, interchange letters p and q. This gives

$$\frac{\Sigma q_k p_k}{\Sigma p_k} \div \frac{\Sigma q_0 p_0}{\Sigma p_0}.$$

The product of the two results,

$$\frac{\dfrac{\Sigma p_k q_k}{\Sigma q_k}}{\dfrac{\Sigma p_0 q_0}{\Sigma q_0}} \times \frac{\dfrac{\Sigma q_k p_k}{\Sigma p_k}}{\dfrac{\Sigma q_0 p_0}{\Sigma p_0}} = \frac{(\Sigma p_k q_k)^2}{(\Sigma p_0 q_0)^2} \times \frac{\Sigma p_0 \cdot \Sigma q_0}{\Sigma p_k \cdot \Sigma q_k},$$

is seen not to equal the ratio of values,

$$\frac{\Sigma p_k q_k}{\Sigma p_0 q_0}.$$

The factor reversal test is not satisfied.

For the data on page 285, using 1900 as base,

$$w_k = \frac{\Sigma p_k q_k}{\Sigma q_k} = \frac{40 \times 6 + 50 \times 2}{6 + 2} = 42.5.$$

$$w_0 = \frac{\Sigma p_0 q_0}{\Sigma q_0} = \frac{20 \times 3 + 100 \times 4}{3 + 4} = 65.7.$$

$$\frac{w_k}{w_0} = \frac{42.5}{65.7} = 0.65.$$

Applying factor reversal,

$$w_k' = \frac{\Sigma q_k p_k}{\Sigma p_k} = \frac{6 \times 40 + 2 \times 50}{40 + 50} = 3.78.$$

$$w_0' = \frac{\Sigma q_0 p_0}{\Sigma p_0} = \frac{3 \times 20 + 4 \times 100}{20 + 100} = 3.83.$$

$$\frac{w_k'}{w_0'} = \frac{3.78}{3.83} = 0.99.$$

The product of the two results is

$$0.65 \times 0.99 = 0.64.$$

This is considerably smaller than the total value ratio,

$$\frac{\Sigma p_k q_k}{\Sigma p_0 q_0} = \frac{340}{460} = 0.74.$$

The time reversal test applied to Fisher's 3 in the form of 53,

$$\frac{\Sigma p_k q_0}{\Sigma p_0 q_0},$$

(p. 280), gives the following results. With 1900 as base the index number is

$$\frac{40 \times 3 + 50 \times 4}{20 \times 3 + 100 \times 4} = \frac{320}{460} = 0.696 = 70 \text{ per cent.}$$

With 1905 as base,

$$\frac{20 \times 6 + 100 \times 2}{40 \times 6 + 50 \times 2} = \frac{320}{340} = 0.941 = 94 \text{ per cent.}$$

These results say that the average price for 1905 is 70 per cent of that for 1900, and that the average price for 1900 is 94 per cent of that for 1905. Or, in each year the prices are lower than they were the other year, which is absurd. The time reversal test is not satisfied.

By examining index number formulas, it will be found that a number of those in common use will satisfy the time reversal test, while but few satisfy the factor reversal test. The problem, then, is to find a formula that satisfies all tests and is not too complex for practical use.

17. Bias.—Many index numbers have what is known as *bias;* that is, they consistently give results too large, or else too small, to be a fair representation of the facts. The simple arithmetic mean of price relatives has an upward bias.[1] The simple harmonic mean has a downward bias. Also each system of weighting has its bias. Weighting each price relative by the value of the commodity in the base year p_0q_0 gives a downward bias. Weighting each price relative by the value of the commodity for a given year p_kq_k gives an upward bias.

18. Crossing.—Fisher's 3 (p. 280) is an arithmetic type weighted by p_0q_0. Fisher's 19,

$$\frac{\Sigma p_kq_k}{\Sigma p_kq_k\dfrac{p_0}{p_k}},$$

is a harmonic type weighted by p_kq_k. Professor Fisher gives good reasons for using an average of these two. The geometric mean seems to be the fairest average to use, lying between the arithmetic mean and the harmonic mean. Taking the average of two index number formulas is called *crossing* them. The

[1] For a general discussion of bias see PROF. IRVING FISHER, "Making of

geometric cross, between Fisher's 3 and 19, is called by him the "Ideal" index number. It is

$$\sqrt{\frac{\Sigma p_0 q_0 \frac{p_k}{p_0}}{\Sigma p_0 q_0} \times \frac{\Sigma p_k q_k}{\Sigma p_k q_k \frac{p_0}{p_k}}}.$$

This reduces to

$$\sqrt{\frac{\Sigma p_k q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_k q_k}{\Sigma p_0 q_k}}.$$

In this latter form the first fraction is ratio of values, the value of each separate commodity being weighted by the quantity of that commodity produced in the base year. The second fraction is the same, except that the weights are the quantities produced in year k. The first fraction is Fisher's 53, and the second, Fisher's 59, two separate index number formulas.

19. Fisher's "Ideal" Satisfies the Tests.—This formula, Fisher's Ideal, satisfies all three tests. That it satisfies the time reversal test may be seen by putting k for every 0, and 0 for every k in the formula and multiplying the result by the original formula.

$$\sqrt{\left(\frac{\Sigma p_0 q_k}{\Sigma p_k q_k} \times \frac{\Sigma p_0 q_0}{\Sigma p_k q_0}\right)\left(\frac{\Sigma p_k q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_k q_k}{\Sigma p_0 q_k}\right)} = 1.$$

The quantities under the radical cancel out, giving $\sqrt{1} = 1$, and the test is satisfied.

To see that the factor reversal test is satisfied, write p for each q, and q for each p. Multiply the result by the original formula and see if it gives the ratio of total values. This process gives

$$\sqrt{\left(\frac{\Sigma q_k p_0}{\Sigma q_0 p_0} \times \frac{\Sigma q_k p_k}{\Sigma q_0 p_k}\right)\left(\frac{\Sigma p_k q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_k q_k}{\Sigma p_0 q_k}\right)} = \sqrt{\frac{(\Sigma p_k q_k)^2}{(\Sigma p_0 q_0)^2}} = \frac{\Sigma p_k q_k}{\Sigma p_0 q_0},$$

the ratio of total values. So the test is satisfied.

20. Cross-weighting.—Another very good formula is obtained by weighting each p by the arithmetic cross of the weights q_0 and q_k. This gives

$$\frac{\sum \frac{q_0 + q_k}{2} p_k}{\sum \frac{q_0 + q_k}{2} p_0}.$$

The common factor ½ to every term of the summations cancels out giving

$$\frac{\Sigma(q_0 + q_k)p_k}{\Sigma(q_0 + q_k)p_0}.$$

The virtue of this formula is that it is easier to compute than the Ideal, and is found empirically to differ from it but little in results.

21. Arbitrary Weights.—Sometimes information as to quantities produced is difficult to obtain. The U. S. census reports give this infomation for every 10 years, but not for every year or oftener. Arbitrary weights may then be assumed, based on estimates. An example of such index number formula is

$$\frac{\Sigma wp_k}{\Sigma wp_0}.$$

This is Fisher's 9,051. *It is of the aggregative type, the price* of each commodity having an arbitrary constant weight. It is the same as Fisher's 53, except that arbitrary weights are used instead of quantities produced in the base year.

22. A Combination That Facilitates Computation.—Professor *Fisher uses a combination of 53 and 9,051 for a weekly index of* 200 commodities. With the 28 most important of the 200 commodities, great care is used in determining the q_0 for each one. For the other 172 commodities, powers of 10 are chosen as arbitrary weights, w, of each. The power of 10 used for each com*modity is that nearest the actual statistical value of the* quantity of that commodity. Some of these 172 commodities have the same weight. Their prices may be added and the decimal point placed according to what power of 10 is used as weight. For each of the 28, a separate product p_kq_0 must be obtained. The result is said to be quite close to that obtained by 53 alone.

23. Computation.—Now calculate Fisher's 53, $\Sigma p_kq_0/\Sigma p_0q_0$, for the data of Tables LXXX and LXXXI. The computations are shown in Table LXXXIII.

The grains are placed in column to facilitate the additions. The column headings are self-explanatory. The figures of the product columns are given to the nearest 100. This will assure accuracy in the results to the nearest 1 per cent.

According to these results the average price of the seven grains listed was, in 1917, 99 per cent greater than in 1913. In 1921 the average price was 77 per cent of that of 1913.

TABLE LXXXIII.—INDEX NUMBERS FOR PRICES OF GRAINS IN THE UNITED STATES FOR YEARS 1917 AND 1921 ON 1913 AS BASE YEAR[1]

Formula, $\frac{\Sigma p_k q_0}{\Sigma p_0 q_0}$

Grain	1913			1917		1921	
	p_0	q_0	$\frac{p_0 q_0}{100}$	p_1	$\frac{p_1 q_0}{100}$	p_2	$\frac{p_2 q_0}{100}$
Corn	69.1	2,447	1,691	127.9	3,129	42.3	1,035
Wheat	79.9	763	610	200.8	1,531	92.6	707
Oats	39.2	1,112	435	66.6	741	30.2	336
Rye	63.4	41	26	166.0	68	69.7	29
Barley	53.7	178	95	113.7	202	41.9	75
Buckwheat	75.5	14	11	160.0	22	81.2	11
Rice	85.8	26	22	189.6	49	95.2	25
Total	....	.....	2,890	.....	5,742	....	2,218

[1] Computations by slide rule.

Index,

$$\frac{2,890}{2,890} \times 100 = 100. \quad \frac{5,742}{2,890} \times 100 = 199. \quad \frac{2,218}{2,890} \times 100 = 77.$$

Table LXXXIV shows the computations for Fisher's Ideal, or 1921 on 1913 as base year.

TABLE LXXXIV.—INDEX NUMBER OF PRICES OF GRAINS IN THE UNITED STATES FOR 1921 ON 1913 BASE[1]

Fisher's Ideal, $\sqrt{\frac{\Sigma p_k q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_k q_k}{\Sigma p_0 q_k}}$

Grain	1913			1921				
	p_0	q_0	$\frac{p_0 q_0}{100}$	p_1	q_1	$\frac{p_1 q_1}{100}$	$\frac{p_1 q_0}{100}$	$\frac{p_0 q_1}{100}$
Corn	69.1	2,447	1,691	42.3	3,069	1,298	1,035	2,121
Wheat	79.9	763	610	92.6	815	755	707	651
Oats	39.2	1,112	435	30.2	1,078	325	336	422
Rye	63.4	41	26	69.7	62	43	29	39
Barley	53.7	178	95	41.9	155	65	75	83
Buckwheat	75.5	14	11	81.2	14	17	11	11
Rice	85.8	26	22	95.2	38	36	25	33
Total	....	.....	2,890	....	.....	2,539	2,218	3,360

[1] Computations by slide r

Index

$$100. \qquad 100 \times \sqrt{\frac{2,218}{2,890} \times \frac{2,539}{3,360}}.$$

The columns for 1913 are the same as in Table LXXXIII. For 1921 two additional columns are necessary, increasing the amount of computation considerably. The computation in the formula is also increased by one division, one multiplication, and a root extraction. The index number turns out to be 76, which is one less than by the other formula. Granting that 76 is more accurate, the question arises as to whether the correction of the error is large enough to compensate for the extra labor. Quite likely the error is no greater than errors in the original data.

The computer can easily make his own table of computations for any index number formula that he desires to use, or for any combination of formulas.

24. Graphs are Valuable.—It is illuminating to graph a series of index numbers over a period of time and extend it periodically, keeping the graph up to date.

25. Warning as to Use of Index Numbers.—Index numbers and their graphs are becoming more and more popular and more diversified in their uses. It is, therefore, important to have an understanding of the principles involved in their making. One should have information in regard to the data used in forming the index number and the formula used in computing it.

One should not apply a wholesale price index to retail prices. A price index based on one set of commodities will likely vary from that based on another set of commodities. Index numbers on prices in one market are likely to vary from those on prices from another market. If the same set of commodities is used, but with different weights, the resulting index numbers should not be compared except with caution. If two index numbers are derived from different bases, the numbers may not be compared without some means of reducing both to the same base. The common method used, if it is desired to reduce index numbers of different bases to a single base, say 1913, is to divide each index of any series by the 1913 index of that series and reduce to per cent. This leads to error in some formulas. For example, the *Annalist's* index of wholesale food prices is based on the average of the decade 1890 to 1899. The index for 1913 is 139.980; for Aug. 2,

1924, it is 184.798. To reduce this latter to 1913 as base, divide 184.798 by 139.980. This gives 132.017. Index numbers should not be compared when derived from different formulas without making allowance for the bias of the formula.

26. An Index Number is a Type Form.—An index number, like any other average, is only a type form. It is the result of an effort to represent somewhat complex and varying data by a single representative type number. While some prices are going up and others going down, while some quantities are increasing and others diminishing, the index cannot be a definite price, but is merely a pointer to show the general trend.

27. Selection of Commodities Used is Important.—*One of the most important prerequisites for a good index number is the selection of the set of commodities used.* A large number of commodities is not so important as the selection of a representative group for the particular purpose at hand. Empirically, it is found that the particular formulas used in most cases differ but little in results. There are some inherently bad formulas, such as the average price relative. The set of weights used is not found to make so much difference in practice as do the commodities selected.

28. Value of Statistical Method.—With index numbers, as with all other statistical work, reliable results, properly interpreted, are to be expected only from those persons who have made themselves familiar with the principles on which statistical method depends. At least an elementary knowledge of the mathematical theory involved is essential to trustworthy results and interpretations.

Exercises

PRODUCTION OF FOOD CROPS IN THE UNITED STATES IN MILLIONS OF BUSHELS

Year	Wheat	Corn	Oats	Barley	Potatoes	Sweet potatoes
1913	763	2,447	1,122	178	332	59
1915	1,025	2,995	1,549	229	360	77
1917	637	3,065	1,593	212	442	84

AVERAGE PRICES RECEIVED BY PRODUCERS IN CENTS PER BUSHEL

Year	Wheat	Corn	Oats	Barley	Potatoes	Sweet potatoes
1913	80	69	39	54	69	73
1915	93	66	45	54	50	79
1917	162	90	51	87	149	90

1. Using the table of food prices of the six food crops given above, find the *average price relative for each year on 1913 as a base year. Apply the time reversal test and see whether the results are consistent with the index* numbers found.

2. Using the table of prices and the table of amounts produced, as given above, find an index number for 1915 and for 1917 on 1913 as base, by means of Fisher's 52, w_k/w_o.

3. *Using the same tables, find an index for 1915 and one for 1917* on *1913* as base, by means of Fisher's 53, $\Sigma p_k q_0 / \Sigma p_0 q_0$.

4. Using the same tables find Fisher's Ideal index for 1915 and for 1917 using 1913 as base.

An abundance of material for problems may be obtained from the large newspaper almanacs. "The Statistical Abstract of the United States" also affords much material. Professor Irving Fisher's "The Making of Index Numbers" has a list of prices and quantities of 36 commodities which he used for his illustrations. Selections from these lists afford very good examples.

APPENDIX A

LOGARITHMS

A logarithm is an exponent. The laws of logarithms are the laws of exponents as developed in algebra.

1. Definition.—The *logarithm* of a number N is that exponent which, operating on a certain specified number, will produce the given number N. The specified number upon which the exponent operates is called the *base* of the system of logarithms. The base of the common or Briggs system of logarithms is 10. The symbol commonly used for logarithm is *log*. The base is indicated as a subscript to the symbol "log." Thus, $\log_{10} 1{,}000 = 3$ is read: The logarithm of 1,000 to the base 10 is 3. The logarithm of 1,000 to the base 10 equals 3, since 10 operated on by the exponent 3 gives 1,000. Thus

$$10^3 = 1{,}000.$$

In general,

$$\log_{10} N = x$$

and

$$10^x = N$$

are equivalent equations stating, in different form, the same relationship between x, N, and 10.

2. Laws of logarithms.—If, now,

$$N = 10^x$$

and

$$M = 10^y,$$

then, by definition of logarithm,

$$\log_{10} N = x$$

and

$$\log_{10} M = y.$$

Moreover, by the laws of exponents,

$$10^x \times 10^y = 10^{x+y}.$$

Therefore,

$$NM = 10^{x+y}$$

and

$$\log_{10} NM = x + y,$$

the exponent of 10 to give NM.

But

$$x = \log_{10} N$$

and

$$y = \log_{10} M;$$

therefore,

$$\log_{10} NM = \log_{10} N + \log_{10} M.$$

This last equation translated into English reads:

(1) *The logarithm of the product of two numbers is the sum their logarithms.*

Similarly,

$$N \div M = 10^x \div 10^y$$
$$= 10^{x-y}.$$

Then,

$$\log_{10} \frac{N}{M} = x - y,$$

or

$$\log_{10} \frac{N}{M} = \log_{10} N - \log_{10} M.$$

This translated into English is:

(2) *The logarithm of the quotient of two numbers is the loǫ rithm of the dividend minus the logarithm of the divisor.*

This same law may be stated:

The logarithm of a fraction is the logarithm of the numera minus the logarithm of the denominator.

Further, since, by the laws of exponents,

$$N^z = (10^x)^z$$
$$= 10^{xz},$$

it follows that

$$\log_{10} N^z = xz$$

or

$$\log_{10} N^z = z \log N.$$

Translated into English this equation reads:

(3) *The logarithm of a power of a number is the index of t power times the logarithm of the number.*

3. Negative, Fractional, and Zero Exponents.—It is proved algebra that the laws of exponents hold whether the exponer are positive or negative, whole numbers or fractions, if prop interpretation is placed in the various kinds of exponents. T following interpretations should be recalled:

$$N^0 = 1, \qquad (a)$$

$$N^{-x} = \frac{1}{N^x}, \qquad (b)$$

$$N^{\frac{p}{q}} = \sqrt[q]{N^p}. \qquad (c)$$

From these interpretations and the laws of logarithms just derived

$$\log_{10} 1 = 0,$$

because

$$10^0 = 1, \qquad (a)$$

$$\log_{10} \frac{1}{N} = \log_{10} N^{-1} \qquad (b)$$

$$= -\log_{10} N, \qquad (3)$$

$$\log_{10} \frac{1}{N^x} = -x \log_{10} N, \qquad (b) \text{ and } (3)$$

and

$$\log_{10} \sqrt[q]{N^p} = \log_{10} N^{\frac{p}{q}}, \qquad (c)$$

$$= \frac{p}{q} \log N. \qquad (3)$$

For example

$$\log_{10} \tfrac{1}{20} = -\log_{10} 20,$$
$$\log_{10} \sqrt{20} = \tfrac{1}{2} \log_{10} 20,$$
$$\log_{10} \sqrt[3]{20^2} = \tfrac{2}{3} \log_{10} 20.$$

4. Base 10.—Since counting goes by powers of 10, the convenient base to use for computation purposes is 10. In ordinary computations where 10 is understood to be the base, it is unnecessary to write the base and only the symbol "log" is used.

The values of the following logarithms to the base 10 are seen at once by inspection:

$$\log 10{,}000 = 4,$$
$$\log 1{,}000 = 3,$$
$$\log 100 = 2,$$
$$\log 10 = 1,$$
$$\log 1 = 0,$$
$$\log 0.1 = -1,$$
$$\log 0.01 = -2,$$
$$\log 0.001 = -3.$$

5. Mantissa and Characteristic.—Now log 20 must be more than 1 and *less than* 2, *since* 20 *lies between* 10 *and* 100. In

other words log 20 is 1 and a fraction. Log N is not an integer unless N is an integral power of 10. If N is not an integral power of 10, then log N is fractional. The decimal part of the value of a logarithm is called the *mantissa* of the logarithm. The part before the decimal point is called the *characteristic*. Tables of logarithms usually give mantissas only. Looking in a six-place table of logarithms the mantissa for 20 is found to be 301,030. Now write

$$\log 20 = 1.301030.$$

Since

$$\begin{aligned} 200 &= 10 \times 20, \\ \log 200 &= \log 10 + \log 20. \qquad (1) \\ &= 1 + 1.301030, \end{aligned}$$

or,

$$\log 200 = 2.301030.$$

Similarly,

$$\log 2{,}000 = 3.301030.$$

Also, since

$$\begin{aligned} 2 &= 20 \div 10, \\ \log 2 &= \log 20 - \log 10 \qquad (2) \\ &= 0.301030. \end{aligned}$$

Moreover,

$$0.2 = 2 \div 10,$$

so

$$\begin{aligned} \log 0.2 &= \log 2 - \log 10 \\ &= 0.301030 - 1. \\ &= 9.301030 - 10 \\ &= -9.698970. \end{aligned}$$

The second form, 9.30103 − 10, is usually the most convenient to use in making computations. It is sometimes written $\bar{1}.301030$. The minus sign being placed over the 1 indicates that it belongs with the 1 alone and not with 301030. This method of indicating a negative logarithm causes inconvenience when several *logarithms are to be added or subtracted and some of them are* negative.

Further, since

$$\begin{aligned} 0.02 &= 0.2 \div 10, \\ \log 0.02 &= \log 0.2 - \log 10 \\ &= 0.301030 - 2 \end{aligned}$$

Similarly,

$$\log 0.002 = 7.301030 - 10.$$

Notice that the mantissa is the same, 301,030, for 2,000, 200, 20, 2, 0.2, 0.02, and 0.002. The number 2 multiplied by any integral power of 10 must have the mantissa 301,030. The characteristic will be the index of the power of 10 used.

Thus:

$$\begin{aligned} 2{,}000 &= 2 \times 10^3 \\ 200 &= 2 \times 10^2, \\ 20 &= 2 \times 10^1, \\ 2 &= 2 \times 10^0, \\ 0.2 &= 2 \times 10^{-1}, \\ 0.02 &= 2 \times 10^{-2}, \\ 0.002 &= 2 \times 10^{-3}, \end{aligned}$$

and so on. The characteristic of the logarithm of

$$\begin{aligned} 2{,}000 &\text{ is } 3, \\ 200 &\text{ is } 2, \\ 20 &\text{ is } 1, \\ 2 &\text{ is } 0, \\ 0.2 &\text{ is } -1, \\ 0.02 &\text{ is } -2, \\ 0.002 &\text{ is } -3, \end{aligned}$$

and so on. Consideration of the above figures will show that, *necessarily, the characteristic of the logarithm of a number greater* than 1 will be one less than the number of digits to the left of the decimal point. The characteristic of the logarithm of a number less than 1 will be negative, and numerically one greater than the number of zeros between the decimal point and the left-hand figure of the number.

6. Rule for the Characteristic.—*These two statements may* be combined into one rule for determining the characteristic of the logarithm of a number.

To determine the *characteristic* of the logarithm of a number, begin at units place and count to that figure, not zero, which is farthest to the left. Count, not 1, 2, 3, etc., but 0, 1, 2, 3, etc. The count upon arriving at the left-hand figure is the characteristic of the logarithm. It is positive if the counting proceeded to the left and negative if the counting proceeded to the right.

Applying this rule, the characteristic of the logarithm of

$$\begin{aligned} 4{,}763.21 &\text{ is } 3, \\ 160.00 &\text{ is } 2, \\ 39.98 &\text{ is } 1, \\ 3.108 &\text{ is } 0, \\ 0.4600 &\text{ is } -1 = 9 - 10, \\ 0.0460 &\text{ is } -2 = 8 - 10, \\ 0.0046 &\text{ is } -3 = 7 - 10, \end{aligned}$$

and so on.

7. Mantissa Independent of the Decimal Point.—Since moving the decimal point in any number merely multiplies or divides the number by an integral power of 10, the logarithm of the number will be changed only by the index of the power of 10 used. That is, since moving the decimal point of a number two places to the right multiplies by 10^2, the logarithm of the number is increased by 2. Or, since moving the decimal point of a number three places to the left divides by 10^3, the logarithm of the number is diminished by 3. It follows that the mantissa for any sequence of figures remains the same no matter where the decimal point is placed.

Characteristics are determined by the position of the decimal point. Mantissas are determined by the sequence of figures. For this reason, tables of logarithms need tabulate mantissas only.

Tables of mantissas are commonly printed to 4, 5, 6, 7, or 8 figures. The greater the number of figures given the greater the degree of accuracy of computation attainable in their use. In general, n-place tables give about n-place accuracy in the computations.

8. Use of the Table.—On pages 349 to 352 is printed a four-place table. To find the logarithm of 46.78, proceed as follows: Write down the characteristic which is seen by inspection to be 2. Follow down the N column of the table to the line where occurs the first two figures, 46, of the number. Follow to the right on this line to the column whose heading is 7, the third figure of the number. Here is found 6,693. This is the mantissa for the sequence of figures 4,670. On the same line in the next column is found 6,702. This is the mantissa for the sequence of figures 4,680. The difference between these two mantissas is

$$6{,}702 - 6{,}693 = 9.$$

Thus for an increase in the number of

$$4{,}680 - 4{,}670 = 10,$$

the mantissa increases 9. Our sequence of figures, 4,678, is 8 more than 4,670. If an increase of 10 in the number causes an increase of 9 in the mantissa, an increase of 8 in the number will give an increase of $\frac{8}{10}$ of 9 in the mantissa, or 7.2. This increase is nearer to 7 than 8; so increase the mantissa by 7, giving

$$6{,}693 + 7 = 6{,}700$$

for the mantissa of the sequence 4,678.

The characteristic 2 has already been determined and written down. Now annex the mantissa, giving

$$\log 46.78 = 2.6700.$$

9. Interpolation.—Finding the mantissa for a number between two numbers as given in the table is called *interpolation*. Inter-

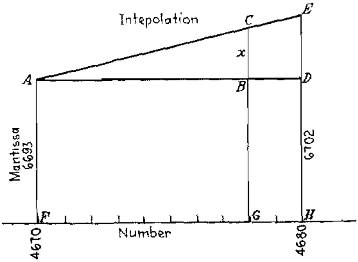


FIG. 66.

polation by the above method assumes uniform variation from one mantissa of the table to the next. It is illustrated in Fig. 66.

Numbers are laid off on the horizontal scale and mantissas on the vertical. $FA = 6{,}693$ is the mantissa for 4,670, and $HE = 6{,}702$ is the mantissa for 4,680. GC, the mantissa for 4,678, is to be found. To do this, find

$$x = BC,$$

the amount to be added to

$$GB = FA = 6{,}693.$$
$$DE = HE - FA = 6{,}702 - 6{,}693 = 9$$

From similar triangles

$$\frac{x}{DE} = \frac{AB}{AD},$$

or

$$\begin{aligned} x &= \frac{AB}{AD} \times DE \\ &= \frac{4{,}678 - 4{,}670}{4{,}680 - 4{,}670} \times 9 \\ &= 8/10 \times 9 \\ &= 7.2. \end{aligned}$$

Then,

$$GC = 6{,}693 + 7 = 6{,}700,$$

the required mantissa.

10. The Inverse Process.—If a logarithm is given as

$$\log N = 9.4823 - 10,$$

find N, the number it is the logarithm of, by use of the table. This number N is called the *antilogarithm* and is written thus:

$$N = \text{antilog } 9.4823 - 10$$

or

$$N = \log^{-1} 9.4823 - 10.$$

To find N, take the mantissa, 4,823, and find the two nearest mantissas, one less and the other greater than 4,823, namely, 4,814 and 4,829; 4,814 is seen to be the mantissa for 3,030; 4,829 is the mantissa for 3,040. The difference of these mantissas is

$$4{,}829 - 4{,}814 = 15.$$

The difference between the given mantissa and the smaller of the two from the table is

$$4{,}823 - 4{,}814 = 9.$$

For a change of 15 in mantissas, there is here a change of 10 in the corresponding number sequence. Then, for a change of 9 in mantissas, there will be a change in number sequence of

$$9/15 \text{ of } 10 \approx 6.$$

The correct number sequence to four figures is, then,

$$3{,}030 + 6 = 3{,}036.$$

Since the characteristic of the given logarithm is $9 - 10 = -1$, the decimal point must be immediately preceding this number sequence and

$$- 10.$$

This process of interpolation is the inverse of that used before. Now GC is given, and FA and HE are found from the table. $FH = 10$ and $FG = AB$ is to be found.

11. Other Bases.—In order to have a system of logarithms, it is not necessary that the base be 10. If 2 is taken as base, then

$$\log_2 8 = 3,$$

since

$$2^3 = 8.$$

Since

$$9^3 = 729 = 27^2,$$

$$\log_9 729 = 3,$$

and

$$\log_{27} 729 = 2.$$

It is evident that there is some relation between the logarithm of a number to one base and its logarithm to another base. To find the relation between $\log_a N$ and $\log_b N$, let

$$\log_a N = x,$$

$$\log_b N = y,$$

and

$$\log_a b = c.$$

Then

$$N = a^x,$$

$$N = b^y,$$

$$b = a^c.$$

Since

$$b = a^c$$

$$b^y = a^{cy}.$$

Now

$$a^x = b^y = N,$$

and

$$a^x = a^{cy}.$$

So

$$x = cy,$$

or

$$\log_a N = \log_a b \cdot \log_b N,$$

whence

$$\log_b N = \frac{\log_a N}{\log_a b}.$$

From this relation, if a table of logarithms to the base a is given, it is possible at once to get logarithms to the base b. Take $N = a$ and substitute a for N in the above relation; then

$$\log_b a = \frac{\log_a a}{\log_a b}.$$

Since $\log_a a$ must equal 1, this gives

$$\log_b a = \frac{1}{\log_a b}.$$

That is, $\log_b a$ and $\log_a b$ are reciprocals of each other. So that

$$\log_b N = \log_a N \cdot \log_b a.$$

12. Natural or Napierian Logarithms.—Besides the common or Briggsian system to the base 10, there is used in theoretical work another system known as the natural, hyperbolic, or Napierian system of logarithms. The base of this system is an incommensurable number indicated by the letter e. To eigbt figures

$$e = 2.7182818.$$

Tables of logarithms are derived by means of certain infinite series which give logarithms to the base e. From logarithms to the base e, logarithms to the base 10 may be found by the relation just determined.

$$\log_{10} N = \log_e N \cdot \log_{10} e.$$
$$\log_{10} e = \log_{10} 2.71828 = 0.434294.$$

Then

$$\log_{10} N = 0.434294 \log_e N.$$

13. Modulus.—This number, 0.434294, is known as the *modulus* of the common system. Its reciprocal, $\log_e 10 = 2.30259$, is the modulus of the natural system.

Having natural logarithms, common logarithms are obtained by multiplying by the modulus of the common system. Having common logarithms, natural logarithms may be obtained by multiplying by the modulus of the natural system. For example, from tables of natural logarithms,

$$\log_e 5 = 1.60944.$$

Then

$$\log_{10} 5 = 0.434294 \times 1.60944$$
$$= 0.698970.$$

From tables of common logarithms,

$$\log_{10} 5 = 0.698970$$

Then

$$\log_e 5 = 2.30259 \times 0.698970$$
$$= 1.60944.$$

14. Cologarithm.—The logarithm of the reciprocal of a number s called the *cologarithm* of the number. Thus

$$\log \frac{1}{N} = colog\ N.$$

From this equation, since

$$\log \frac{1}{N} = \log 1 - \log N$$

$$= 0 - \log N,$$

t follows that

$$\text{colog } N = -\log N.$$

Therefore, whenever in computations log N is to be subracted, it serves the same purpose to add colog N. To find olog N, subtract log N from 0 in the form of 10 − 10. Thus

$$\log 20 = 1.3010.$$

To get colog 20, subtract as follows:

$$\begin{array}{rl} & 10.0000 - 10 \\ & \underline{\ \ 1.3010\qquad\ } \\ \text{colog } 20 & \ \ 8.6990 - 10. \\ \log 0.0341 & = 8.5328 - 10. \end{array}$$

To get colog 0.0341 subtract from 10 − 10, thus:

$$\begin{array}{rl} & 10.0000 - 10 \\ & \underline{\ \ 8.5328 - 10} \\ colog\ 0.0341 & = \ 1.4672 \end{array}$$

It is a simple matter to write down colog N from log N itself. Begin at the characteristic and subtract each figure from 9 except he figure farthest to the right that is not 0. Subtract this last igure from 10. Write down the results of subtraction as you roceed. If log N is positive, −10 must be written after the esult of the subtractions. If log N has −10 written as part of t, there is no −10 with colog N.

15. Examples.—A few illustrative examples of logarithmic computation are given below:

$$x = 245.6 \times 1.391.$$

$$\log x = \log 245.6 + \log 1.391. \qquad (1)$$

$$\begin{aligned} \log 245.6 &= 2.3902 \\ \log 1.391 &= 0.1433 \\ \hline \log x &= 2.5335 \\ x &= 341.6 \end{aligned}$$

If

$$x = 2.492 \div 0.1265.$$

$$\log x = \log 2.492 - \log 0.1265. \qquad (2)$$

$$\begin{aligned} \log 2.492 &= 0.3965 \\ \log 0.1265 &= 9.1021 - 10 \\ \hline \log x &= 1.2944 \\ x &= 19.70 \end{aligned}$$

The table gives the mantissa for 1,970 as 2,945; 2,944 is so close to 2,945 that the interpolation gives a sequence of figures closer to 1,970 than to 1,969.

This division may be performed by the equation;

$$\log x = \log 2.492 + \text{colog } 0.1265.$$

$$\begin{aligned} \log 2.492 &= 0.3965 \\ \text{colog } 0.1265 &= 0.8979 \\ \hline \log x &= 1.2944 \\ x &= 19.70 \end{aligned}$$

If

$$x = \frac{0.0345\sqrt{64.8}}{0.00374\sqrt[3]{9.34}},$$

$$\log x = \log 0.0345 + \tfrac{1}{2} \log 64.8 + \text{colog } 0.00374 + \tfrac{1}{3} \text{ colog } 9.34.$$

$$\begin{aligned} \log 0.0345 &= 8.5378 - 10 \\ \tfrac{1}{2} \log 64.8 &= 0.9058 \\ \text{colog } 0.00374 &= 2.4271 \\ \tfrac{1}{3} \text{ colog } 9.34 &= 9.6766 - 10 \\ \hline \log x &= 1.5473 \\ x &= 35.3. \end{aligned}$$

Facility in the use of logarithms is attained only by practice.

APPENDIX B

PERMUTATIONS AND COMBINATIONS. BINOMIAL EXPANSION

If a group of r things is selected from a set of n things, this group of r things is called a *combination*. The question at once arises as to how many different groups of r things it is possible to select from n things. In other words, it is desired to know the *number of combinations* of n things taken r at a time. The symbol for this number of combinations is ${}_nC_r$.

In any set or combination of r things, the things may be arranged in different orders. Each arrangement of the r things is called a *permutation* of the r things. The symbol for the *number of permutations* of n things taken r in each set is ${}_nP_r$.

From the four letters a, b, c, d, may be selected the following sets of three each, *abc*, *abd*, *acd*, and *bcd*. There being four and only four of these possible sets, ${}_4C_3 = 4$.

In any one of these sets or combinations of letters, such as *abc*, for example, the three letters may take any one of the following arrangements, *abc*, *acb*, *bac*, *bca*, *cab*, or *cba*, six in number. Each of these arrangements is a permutation of the three letters. The three letters in each of the four combinations have six permutations. Then for all four combinations there are $4 \times 6 = 24$ different permutations. So ${}_4P_3 = 24$. To find the value of ${}_nC_r$ and ${}_nP_r$, first consider the following theorem.

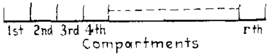
1. Fundamental Theorem.—*If there are p ways of performing one act and q ways of performing an act associated with it, then there are pq ways in which the two associated acts may be performed.*

This is shown by the following illustration. If there are five roads from New York to Chicago and four roads from Chicago to Seattle, then there are $5 \times 4 = 20$ routes from New York to Seattle by way of Chicago. This is true since there are five choices of road from New York to Chicago, and with any one of these five may be associated any one of four roads from Chicago to Seattle, making 5×4 ways to travel from New York to Seattle by way of Chicago. Similarly, if there are p roads from New York

to Chicago and q roads from Chicago to Seattle, there are p choices of routes from New York to Chicago, and with the one chosen may be associated any one of q roads to finish the trip, making pq choices of route from New York to Seattle by way of Chicago.

If, in addition, there are three roads from Seattle to San Francisco, then any one of the 20 ways from New York to Seattle may be associated with any of the three ways to San Francisco, resulting in $20 \times 3 = 5 \times 4 \times 3 = 60$ choices of route from New York to San Francisco by way of Chicago and Seattle.

2. Value of $_nP_r$.—In order to find the value of $_nP_r$, suppose there are r compartments to be filled by selections of r things from n things, placing one thing in each compartment. (See figure.)



The first compartment may be filled with any one of the n things. That is, there are n choices of the thing with which to fill the first compartment. When this compartment is filled, there are $n - 1$ things left from which to select one to fill the next compartment. So, by the theorem just stated, the number of ways in which the first and second compartments may be filled is $n(n - 1)$. With two compartments filled, there are $n - 2$ things left from which to select one to fill the third compartment. The three compartments may then be filled in $n(n - 1)(n - 2)$ ways. Similarly, the first four may be filled in $n(n - 1)(n - 2)(n - 3)$ ways. It is at once seen that the number of ways the r compartments may be filled is the continued product of $n(n - 1)(n - 2) \ldots (n - r + 1)$. The number of ways in which the r compartments may be filled is $_nP_r$. So $_nP_r = n(n - 1)(n - 2) \ldots (n - r + 1)$. It follows at once that the number of permutations of n things taken all, or n, at a time is $n(n - 1)(n - 2) \ldots 3 \cdot 2 \cdot 1$. This product of the natural numbers from 1 to n inclusive is called factorial n. The symbol for factorial n is $n!$ or $\underline{|n}$.

$$\underline{|5} = 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 = 120.$$

In the product $n(n-1)(n-2) \ldots (n-r+1)$ there are all of the factors in $\underline{|n}$ except $1 \cdot 2 \cdot 3 \ldots (n-r) = \underline{|n-r}$.

Therefore,

$$_nP_r = \frac{\underline{|n}}{\underline{|n-r}}.$$

From this formula,

$$_4P_3 = \frac{\underline{|4}}{\underline{|1}} = \frac{1 \cdot 2 \cdot 3 \cdot 4}{1} = 24,$$

as was previously stated.

3. Value of $_nC_r$.—If a combination of r things be selected, these r things have $\underline{|r}$ permutations among themselves, since $_rP_r = \underline{|r}$. Each of the combinations, then, of n things taken r at a time, has $\underline{|r}$ permutations among the things in the combination. It follows, at once, that

$$_nC_r \cdot \underline{|r} = {}_nP_r.$$

Therefore,

$$_nC_r = \frac{_nP_r}{\underline{|r}},$$

or

$$_nC_r = \frac{\underline{|n}}{\underline{|n-r}\,\underline{|r}}.$$

From this formula,

$$_4C_3 = \frac{\underline{|4}}{\underline{|1}\,\underline{|3}} = 4,$$

as was previously obtained.

If, from a set of n things, all n of them are selected, there is, of course, only one such combination. That is $_nC_n = 1$. By the formula for number of combinations, $_nC_n = \frac{\underline{|n}}{\underline{|n-n}\,\underline{|n}}$. Since this must equal 1, the formula is consistent only by assuming $\underline{|n-n} = 1$. That is, $\underline{|0} = 1$.

4. A Simple Relationship.—$_nC_r = {}_nC_{n-r}$. This is seen by applying the formula to each member of the equation.

$$_nC_r = \frac{\underline{|n}}{\underline{|n-r}\,\underline{|r}}.$$

$$_nC_{n-r} = \frac{\underline{|n}}{\underline{|n-(n-r)}\,\underline{|n-r}} = \frac{\underline{|n}}{\underline{|r}\,\underline{|n-r}}.$$

This last result is the same as the first; therefore, ${}_nC_r = {}_nC_{n-r}$ This is an important theorem.

5. Binomial Expansion.—When $(a + b)^n$ is multiplied out the *resulting series is known as the binomial series* and the process of getting the series is the *binomial expansion.* Such simple cases as

$$(a + b)^2 = a^2 + 2ab + b^2$$

and

$$(a + b)^3 = a^3 + 3a^2b + 3ab^2 + b^3$$

are well known in elementary algebra. In multiplying out $(a + b)^n$, the expression becomes $(a + b)(a + b)(a + b)$ . . . to n factors. To write down the result, all combinations of products that can be obtained by taking one letter from each factor must be obtained and all these products added together. Taking a from each factor gives a^n. Taking a from each factor except the first gives $a^{n-1}b$. Taking a from each factor except the second also gives $a^{n-1}b$. Similarly, taking a from each factor but one gives $a^{n-1}b$ as many times as there are factors, or n times The sum of these products is $na^{n-1}b$. A simple manner of looking at this process is to consider it as taking $a^{n-1}b$ as many times as the number of times it is possible to select one thing from n things, that is ${}_nC_1$. This is true since b is selected from one factor in as many ways as there are factors. Next consider in how many ways a may be selected from all but two factors and b be selected from the other two. This must be ${}_nC_2$. Therefore in the expansion will occur ${}_nC_2a^{n-2}b^2$.

Similarly, the number of ways a may be selected from all but three factors and b be selected from the remaining three is ${}_nC_3$ and the term ${}_nC_3a^{n-3}b^3$ occurs in the expansion. For a selected from all but four factors and b from the remaining four, the term ${}_nC_4a^{n-4}b^4$ occurs. This process will be continued to the point where a is selected from one factor and b from all the rest, giving the term ${}_nC_{n-1}ab^{n-1}$. Only one more term will follow, namely that obtained by selecting b from each factor, giving the last term, b^n. The terms that are thus obtained being added together gives the series:

$$(a + b)^n = a^n + {}_nC_1a^{n-1}b + {}_nC_2a^{n-2}b^2 + {}_nC_3a^{n-3}b^3 + {}_nC_4a^{n-4}b^4 + \cdots + {}_nC_{n-1}ab^{n-1} + b^n$$

Since ${}_nC_r = {}_nC_{n-r}$, it follows that the coefficient of the second term equals the coefficient of next to the last term. ${}_nC_1 = {}_nC_{n-1}$ Also the coefficient of the third term equals that of the third

erm from the end. ${}_nC_2 = {}_nC_{n-2}$. The same relation holds hroughout the series. Terms equally distant from the beginning ,nd end of the series have equal coefficients. Evidently, if there s an even number of terms, the coefficients will all be paired. f there is an odd number of terms, there will be a middle term vhose coefficient is not paired. From the manner in which the ight-hand subscripts run in the coefficients, it is evident that here are $n + 1$ terms in the expansion. This is also evident rom the manner in which the exponents of b run, as well as those of a also. In the series as written above, it may be noticed that the right-hand subscript and the exponent of b are always the same. From the theorem in combinations just used, the right-hand subscript for the coefficient of any term may just as well be made the same as the exponent of a in that term.

By writing down the values of these coefficients in terms of n, a law of formation may be found.

$${}_nC_1 = {}_nC_{n-1} = \frac{\underline{|n}}{\underline{|n-1}\,\underline{|1}} = n,$$

$${}_nC_2 = {}_nC_{n-2} = \frac{\underline{|n}}{\underline{|n-2}\,\underline{|2}} = \frac{n(n-1)}{\underline{|2}} = \frac{n(n-1)}{1\cdot 2},$$

$${}_nC_3 = {}_nC_{n-3} = \frac{\underline{|n}}{\underline{|n-3}\,\underline{|3}} = \frac{n(n-1)(n-2)}{\underline{|3}} = \frac{n(n-1)(n-2)}{1\cdot 2\cdot 3},$$

$${}_nC_4 = {}_nC_{n-4} = \frac{\underline{|n}}{\underline{|n-4}\,\underline{|4}} = \frac{n(n-1)(n-2)(n-3)}{\underline{|4}} = \frac{n(n-1)(n-2)(n-3)}{1\cdot 2\cdot 3\cdot 4},$$

$${}_nC_r = {}_nC_{n-r} = \frac{\underline{|n}}{\underline{|n-r}\,\underline{|r}} = \frac{n(n-1)(n-2)\ .\ .\ .\ (n-r+1)}{\underline{|r}},$$

and so on.

It may be seen, at once, that any coefficient may be obtained by multiplying the numerator of the preceding coefficient by a factor one less than the last factor in that numerator and the denominator by a factor one greater than the last factor in that denominator. Thus,

$${}_nC_3 = {}_nC_2\frac{n-2}{3},$$

and

$${}_nC_4 = {}_nC_3\frac{n-3}{4},$$

and so on.

The binomial series written out with the coefficients in terms of n, gives:

$$(a + b)^n = a^n + na^{n-1}b + \frac{n(n-1)}{1 \cdot 2} a^{n-2}b^2 + \frac{n(n-1)(n-2)}{1 \cdot 2 \cdot 3} a^{n-3}b^3 + \ . \ . \ . \ + nab^{n-1} + b^n.$$

6. Rules for Forming Terms.—From the rule given above for forming any coefficient from the preceding coefficient, and remembering that in any term the exponent of b equals the last factor in the denominator of the coefficient, and that the exponent of a is always n minus the exponent of b, the following rule may be stated for deriving any coefficient from the preceding coefficient:

For the coefficient of any term, multiply the coefficient of the preceding term by the exponent of a in that term and divide by one more than the exponent of b.

The exponent of b in the first term is 0, in the second term it is 1, in the third term 2, and so on. In general, the exponent of b in any term is one less than the number of that term as it stands in the series. The exponent of b in the $(r + 1)$th term is then r. The exponent of a is $n - r$. The last factor of the denominator of the coefficient is r. The last factor of the numerator is $n - r + 1$. Now write the $(r + 1)$th term. It is

$$\frac{n(n-1)(n-2) \ . \ . \ . \ (n-r+1)}{\underline{|r}} a^{n-r}b^r.$$

The $(r + 1)$th term may also be written as ${}_nC_r a^{n-r}b^r$.

7. Pascal's Triangle.—It is of interest to note that the coefficients of the binomial expansion for any value of n may be developed in triangular form.

```
                  1
                1   1
              1   2   1
            1   3   3   1
          1   4   6   4   1
        1   5  10  10   5   1
      1   6  15  20  15   6   1
      . . . . . . . . . . . . . . .
```

The first row, 1, 1, gives the coefficients of $(a + b)^1 = a + b$. The second row, 1, 2, 1, gives the coefficients of $(a + b)^2 = a^2 + 2ab + b^2$, and so on.

In this triangle, any number is equal to the sum of the two numbers just above it, one to the right and the other to the left. For example 6 in the fourth row equals 3 + 3, 10 in the fifth row equals 4 + 6, 1 in any row equals 1 + 0. To form additional rows start with 1 and proceed by adding the numbers two at a time, of the row last written. Thus the next row will be 1, 1 + 6, 6 + 15, 15 + 20, 20 + 15, 15 + 6, 6 + 1, 1 + 0, or 1, 7, 21, 35, 35, 21, 7, 1. This set forms the coefficients in the expansion of

$$(a + b)^7 = a^7 + 7a^6b + 21a^5b^2 + 35a^4b^3 + 35a^3b^4 + 21a^2b^5 + 7ab^6 + b^7.$$

This triangular scheme is known as *Pascal's triangle*.

8. If *n* is Not a Positive Integer.—*If n* is not a positive integer, the laws for binomial expansion may still be followed, but there will be no end to the number of terms in the binomial series. For example, the laws of binomial expansion give

$$(a + b)^{\frac{1}{2}} = a^{\frac{1}{2}} + \frac{1}{2}a^{-\frac{1}{2}}b + \frac{(\frac{1}{2})(-\frac{1}{2})}{\underline{|2}}a^{-\frac{3}{2}}b^2 + \frac{(\frac{1}{2})(-\frac{1}{2})(-\frac{3}{2})}{\underline{|3}}a^{-\frac{5}{2}}b^3 + \cdots$$

It is seen, at once, that the exponent of a will never become zero, permitting the expansion to stop. The infinite series resulting when n is not a positive integer have valuable applications when certain conditions of convergence are fulfilled. These conditions of convergence together with the proof of the binomial expansion when n is not a positive integer are in the province of higher algebra, and will not be presented here. It is sufficient to state that the conditions of convergence are satisfied in the expansion of $(1 + x)^n$, if x is numerically less than one.

For example, expanding $(1 - x)^{-1}$ by the binomial expansion gives

$$(1 - x)^{-1} = 1^{-1} + (-1)1^{-2}(-x) + \frac{(-1)(-2)}{\underline{|2}}1^{-3}(-x)^2 + \frac{(-1)(-2)(-3)}{\underline{|3}}1^{-4}(-x)^3 + \cdots$$

$$= 1 + x + x^2 + x^3 + \cdots$$

This series never ends, or is an infinite series.

Since

$$(1 - x)^{-1} = \frac{1}{1 - x},$$

this same series may be obtained by actual division of 1 by $1 - x$.

In this series, it is evident that if $x \geq 1$, the sum of the series may be made as large as desired by carrying the expansion out to a sufficiently large number of terms.

If $x < 1$, the series is convergent. By convergent is meant that the sum approaches a definite limit as the number of terms is made to increase without limit. It is the well-known geometric progression in which the constant ratio is x.

APPENDIX C

LAWS OF PROBABILITY

If there are several equally likely events that may happen, the probability that a given one of these events will happen is the ratio of the number of cases favorable to its happening to the total number of possible cases.

For example, if a bag contains four white balls and three black ones, all alike except for color, the probability of drawing at random a white ball is $4/7$. There are seven possible events, each the drawing of a ball. Since the balls are alike except for color, these events are equally likely. Of these seven events, four are favorable to drawing a white ball. By the definition, then, the probability of drawing a white ball is $4/7$.

This means that if a ball is drawn and a record is made of whether it is white or black, and then the ball is returned to the sack and another drawing is made, the color of the ball recorded, and the process repeated a great many times, a white ball may be expected to be drawn about four times out of seven. In other words, in the long run about $4/7$, 57 per cent, of the drawings are expected to give white balls. It is necessary, of course, that there be no greater likelihood of drawing any one ball than any other.

If the event does not happen, it fails. In the above illustration the probability of not drawing a white ball is $3/7$. The probability of happening plus the probability of failure equals unity, since the number of favorable cases plus the number of unfavorable cases equals the total number of possible cases.

Let p be the probability of happening and q be the probability of failure; then

$$p + q = 1.$$

Since the event must either happen or fail, the sum of p and q represents certainty.

Theorem I.—The probability that all of a set of independent events will occur is the product of the separate probabilities of each of the single events. This is the *compound probability* of the simulta-

neous happening of all the events or of their happening i succession.

Proof.—Let p_1 be the probability of the happening of the firs event, p_2 that of the second, p_3 of the third, and so on to p_n, th probability of the happening of the nth event. Let the number o possible cases in connection with the first event be m_1 and let a_1 o these be favorable to the happening of that event. Similarly m_2 and a_2 are the number of possible and of favorable cases respec tively to the happening of the second event, m_3 and a_3 for th third event, and so on to m_n and a_n for the nth event. Then, b definition,

$$p_1 = \frac{a_1}{m_1},\ p_2 = \frac{a_2}{m_2},\ p_3 = \frac{a_3}{m_3},\ \ldots\ p_n = \frac{a_n}{m_n}.$$

Any one of the m_1 ways for the first event may be combine with any one of the m_2 ways for the second event, and any one o these combinations may be combined with any one of the m ways for the third event, and so on. Then by the first theorer proved in Appendix B, the number of possible ways for all event is the product $m_1m_2m_3 \ldots m_n$. Any one of the a_1 favorabl cases for the first event may be combined with any one of the a favorable cases for the second event and any one of these combina tions may be combined with any one of the a_3, favorable case for the third event, and so on. So, by the same theorem, th total number of favorable cases for all the events is the produc $a_1\,a_2\,a_3 \ldots a_n$.

The probability p of the simultaneous or successive happenin of all the events is the ratio of the total number of favorable case to the total number of possible cases. Therefore,

$$p = \frac{a_1a_2a_3 \ldots a_n}{m_1m_2m_3 \ldots m_n}.$$

But this fraction equals the product

$$\frac{a_1}{m_1} \cdot \frac{a_2}{m_2} \cdot \frac{a_3}{m_3} \cdot \ldots \cdot \frac{a_n}{m_n}.$$

This product is the product of the separate probabilities of eac of the single events. Therefore,

$$p = p_1p_2p_3 \ldots p_n.$$

The probability of failure of each of these events is

$$(1 - p_1),\ (1 - p_2),\ (1 - p_3),\ \ldots\ (1 - p_n).$$

Corollary I.—The probability of the failure of all n events is the continued product,

$$(1 - p_1)(1 - p_2), (1 - p_3) \; . \; . \; . \; (1 - p_n).$$

Corollary II.—The probability of the happening of r of the n events and the failure of all the rest is

$$p_1p_2p_3 \; . \; . \; . \; p_r(1 - p_{r+1})(1 - p_{r+2}) \; . \; . \; . \; (1 - p_n).$$

If one sack contains four white balls and three black ones, and another sack contains seven white balls and five black, and a ball is drawn from each sack, what is the probability that both balls will be white?.

Drawing a ball from one sack has nothing to do with what color will be drawn from the other sack. So the two drawings (events) are independent.

$$p_1 = \frac{a_1}{m_1} = \frac{4}{7}.$$

$$p_2 = \frac{a_2}{m_2} = \frac{7}{12}.$$

$$p = p_1p_2 = 4/7 \cdot 7/12 = 1/3.$$

So the probability that both balls will be white is $1/3$. This means that if a large number of pairs of drawings were performed, about one-third of them would be expected to yield two white balls.

Theorem II.—The probability that **any one whatever** *of a set of* **mutually exclusive** *events will occur and all the others fail is the sum of the partial probabilities of the single events.*

Proof.—First make clear the meaning of *partial probability.* Suppose a sack contains 27 balls, of which 7 are white, 3 are black, 8 are red, and 9 are green. Let 4 white balls be marked with an X, 2 black ones marked with an X, 3 red ones marked with an X, and 7 green ones marked with an X.

In drawing a ball with an X on it, any one of four *mutually exclusive* events may happen, namely, drawing a white ball with an X, a black ball with an X, a red ball with an X, or a green ball with an X. If a ball is drawn at random from the sack, the probability that some given one of these four events will happen is the *partial probability* of the happening of that particular event.

To prove the theorem, suppose there are n possible happenings and that a_1 is the number of cases favorable to the happening of the first of m mutually exclusive events, a_2 the number favorable to the happening of the second, and so on to a_m, the number

favorable to the happening of the mth of the mutually exclusive events. Also let $p_1, p_2, \ldots p_m$ be the probability of the happening of these events, respectively. Then, by definition of probability,

$$p_1 = \frac{a_1}{n}, \quad p_2 = \frac{a_2}{n}, \quad \ldots \quad p_m = \frac{a_m}{n}.$$

These are partial probabilities. Now, the number of cases favorable to the happening of any one whatever of the m events is

$$a_1 + a_2 + \ldots + a_m.$$

Therefore, the probability of the happening of any event whatever of the series is

$$p = \frac{a_1 + a_2 + \ldots + a_m}{n}.$$

Or

$$p = \frac{a_1}{n} + \frac{a_2}{n} + \ldots + \frac{a_m}{n}$$
$$= p_1 + p_2 + \ldots + p_m,$$

which was to be proved. The sum of the partial probabilities is called *total probability*.

In the illustration of 27 balls in a sack, there are four mutually exclusive events.

$$m = 4.$$

There are $n = 27$ possible happenings. The number favorable to the first event, the drawing of a white ball with an X, is $a_1 = 4$. The number favorable to the second event, drawing a black ball with an X, is $a_2 = 2$. The number favorable to the third event is $a_3 = 3$, and to the fourth event is $a_4 = 7$. The partial probabilities of each of the events are

$$p_1 = 4/27,$$
$$p_2 = 2/27,$$
$$p_3 = 3/27,$$
$$p_4 = 7/27,$$

respectively.

The probability of any event whatever of the set, that is, the probability of drawing a ball with an X is

$$p = 4/27 + 2/27 + 3/27 + 7/27 = 16/27.$$

This result is evident since, of the 27 balls, 16 are marked with an X, and the probability of drawing a ball within an X is $16/27$.

Theorem III.—If the probability that an event will occur in a single trial is p, and the probability it will fail is q, then the probability that it will occur exactly r times, no more no less, in n trials is

$$_nC_r p^r q^{n-r}.$$

Proof.—The probability the event will happen in a given trial is p. Its happening or not happening has nothing to do with its happening in another trial. The probability of its happening in another given trial is also p. Then, by Theorem I, the probability of its happening in any two given trials is the compound probability $p \cdot p = p^2$. By the same theorem, the probability of its failing to happen in all the rest of the n trials is q^{n-2}. Also, by the same theorem, the probability of the simultaneous occurrence of happening in any two given trials and failing in all the rest is

$$p^2q^{n-2}.$$

Similarly, the probability of the simultaneous happening of the event in any three given trials and failure in all the rest is

$$p^3q^{n-3}.$$

In general, the probability of the happening of the event in any r given trials and failure in all the rest is

$$p^rq^{n-r}.$$

The number of ways in which the r given trials may be chosen is

$$_nC_r.$$

The probability of the happening of any one of these combinations of r successes and $n - r$ failures is

$$p^rq^{n-r}.$$

When any one of these combinations occurs, it excludes the happening of any other of the combinations. So, by Theorem II, the probability of the happening of any one whatever of the set of $_nC_r$ events is $p^rq^{n-r} + p^rq^{n-r} + \ldots$ to $_nC_r$ terms.

This sum equals

$$_nC_r p^r q^{n-r},$$

the total probability of the happening of the event exactly r times, no more no less.

Theorem IV.—If the probability that an event will occur in a single trial is p, and the probability that it will fail is q, then the probability that it will occur at least r times in n trials (r times or more) is the sum of the first (n − r + 1) terms in the expansion of (p + q)ⁿ; that is,

$$p^n + {}_nC_1p^{n-1}q + {}_nC_2p^{n-2}q^2 + \ldots + {}_nC_rp^rq^{n-r}.$$

Proof.—In order to happen r times or more, it may happen exactly n times, exactly $n - 1$ times, $n - 2$ times, and so on down to $n - (n - r) = r$ times. Now, by Theorem I, the probability that the event will happen n times is

$$p^n.$$

By Theorem III, the probability that it will happen exactly $n - 1$ times is

$$_nC_1p^{n-1}q,$$

the probability that it will happen exactly $n - 2$ times is

$$_nC_2p^{n-2}q^2,$$

and so on until the probability that it will happen exactly r time is

$$_nC_rp^rq^{n-r}.$$

If the event happens n times in n trials, that excludes its happening exactly $n - 1$ times, $n - 2$ times, or any other number o times. A similar statement is true for exactly $n - 1$ times, and for each of the above series, down to exactly r times. The probability of the happening of any one whatever of this series o events is, by Theorem II, the sum of the separate probabilities o each event. This sum is

$$p^n + {}_nC_1p^{n-1}q + {}_nC_2p^{n-2}q^2 + \quad . \quad . \quad . \quad + {}_nC_rp^rq^{n-r},$$

the probability sought.

The application of these theorems to point binomial frequency distributions has been seen in the body of the text

APPENDIX D

DERIVATIVES AND INTEGRALS

1. Limit of a Variable.—To understand the meaning of "derivative," a clear idea of "limit of a variable" should be obtained. *A variable x is said to approach a constant L as its limit if, as x varies under some law, the value of the difference between x and L can be made to become and remain less than any assignable number, no matter how small.* The fact that L is the limit of x may be written $x \rightarrow L$. This symbol is read "x approaches L as a limit."

2. As an illustration of a variable and its limit, suppose a weight hanging at rest on a coiled wire spring. If the weight is drawn down and let go, the length of the supporting spring changes, getting shorter, then longer, alternately, until the weight finally comes to rest again. While the weight oscillates up and down, the length of the spring is a variable, x, sometimes increasing, sometimes decreasing. The amplitude of oscillation diminishes according to a law governing springs in such a situation. After a time the weight comes to rest, and the spring is at its original length L. A situation has arisen such that the difference between x and L has become zero, and so is less than any assignable quantity, no matter how small. As the difference between x and L will now remain zero, it remains less than any assignable quantity, no matter how small. So by definition L is the limit of x. It is to be noted that at one point in each oscillation, either up or down, the value of x is equal to L, but the difference between x and L does not remain zero. This variable finally reaches its limit and ceases to vary.

3. Another illustration of a variable approaching a limit is given by the sum of the terms of a geometrical progression having a proper fraction as the constant ratio, when the number of terms may be made as large as we please. Let S_n be the sum of the first n terms of the series formed by starting with 1 and forming each term by taking half of the preceding term. Thus:

$$S_4 = 1 + \tfrac{1}{2} + \tfrac{1}{4} + \tfrac{1}{8}.$$
$$S_6 = 1 + \tfrac{1}{2} + \tfrac{1}{4} + \tfrac{1}{8} + \tfrac{1}{16} + \tfrac{1}{32}.$$
$$S_n = 1 + \frac{1}{2} + \frac{1}{4} + \ . \ . \ . \ + \frac{1}{2^{n-1}}.$$

As n changes, S_n varies. n may be made as large as we please. The limit of S_n, as n increases without limit, is 2. This is written $\lim_{n \to \infty} S_n = 2$.

Or, $S_n \to 2$ as $n \to \infty$. That 2 is the limit may be seen in the following manner. Suppose a stick 2 ft. long. Cut it in two in the middle, making two pieces each 1 ft. long. Take half of one of these pieces and add it to the other. These two pieces together make a length of

$$1 + \tfrac{1}{2} = 1\tfrac{1}{2}.$$

Cut the remaining piece in two in the middle and add one of the pieces to the two just added. This makes a total length of

$$1 + \tfrac{1}{2} + \tfrac{1}{4} = 1\tfrac{3}{4}.$$

Repeat the process with the remaining piece, adding half of it to the lengths already added, giving

$$1 + \tfrac{1}{2} + \tfrac{1}{4} + \tfrac{1}{8} = 1\tfrac{7}{8}.$$

The sum always is less than 2 by the length of the piece left over. The lengths of the pieces left over are, successively 1, $\tfrac{1}{2}$, $\tfrac{1}{2^2}$, $\tfrac{1}{2^3}$, and so on. At the $(n + 1)$th division, the length of the piece left over is $\tfrac{1}{2^n}$, and the pieces added together make

$$1 + \tfrac{1}{2} + \tfrac{1}{4} + \ . \ . \ . \ + \tfrac{1}{2^n}.$$

Now, by making n large enough, $\tfrac{1}{2^n}$ may be made as small as we please, though never zero. Since the sum of the pieces added together differs from 2 by the length of the piece left over, this difference may be made as small as we please by making n large enough. Suppose, then, a small quantity is named. The difference between 2 and the sum of the pieces added together may be made less than this small quantity simply by making n large enough. From that point on, as the process of adding half the piece left over is continued, the difference between 2 and the pieces added together will remain less than the small quantity that was named. This can be done no matter how small a quantity may be named. So the process enables us to make the sum of the pieces added together become and remain different from 2 by less than any assignable number, no matter how small. Therefore, by definition, 2 is the limit of the sum of the terms

$$1 + \tfrac{1}{2} + \tfrac{1}{4} + \ . \ . \ . \ + \tfrac{1}{2^n}$$

as n is made to increase without limit.

It is to be noted that there is always a piece left over and the variable sum will never actually reach its limit of 2. In this

respect this variable differs from the variable length of the spring in the first illustration.

4. The variable length of spring was alternately larger and smaller than the limit value. During the process of variation, the length became the same as the limit a number of times. The sum of the lengths of pieces of stick is always less than the limit. This variable never can reach the limit value. A variable might always be greater than its limit. It may reach the limit value or it may not, according to the nature of the variable. The important consideration is that the process of variation may be continued till the difference between the variable and its limit becomes smaller than any specified quantity, no matter how small; and that as the process of variation continues, this difference remains less than the specified quantity, no matter how small that quantity may be.

5. Tangent to a Curve.—In Fig. 67, consider the two points P and Q on a curve, and the secant line PQ through these points.

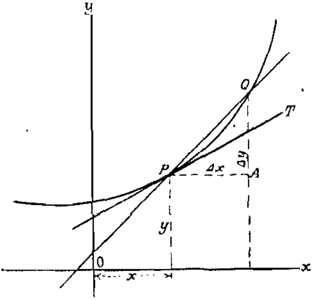


FIG. 67.—Tangent to a curve. Derivative.

Let the coordinates of P be x and y. In going from the point P to the point Q, the x-coordinate changes by an amount $PA = \Delta x$. (Δx is called an *increment* of x. It is read "delta x" or "increment of x." Δ is a symbol indicating "increment of," and is not a quantity multiplying x.) The y-coordinate changes by an amount $AQ = \Delta y$ (increment of y). So the coordinates of Q are $x + \Delta x$ and $y + \Delta y$. The slope of the secant PQ is

$$\frac{AQ}{PA} = \frac{\Delta y}{\Delta x}.$$

6. Now let Q move *along the curve* toward P, keeping the secant line through P and Q as Q moves. Δx and Δy grow smaller and smaller. The secant line turns on the point P. As it turns, its slope is always equal to $\Delta y/\Delta x$. Since the slope is changing, Δx and Δy do not change so as to maintain a constant ratio.

7. Let Q move *along the curve* toward P as a limiting position. This means that Q moves in *such a manner* that its distance, on the curve, from P can be made to become and remain less than any assigned quantity no matter how small. One is not concerned whether Q actually reaches P or not. As Q moves along the curve toward P as a limiting position, the secant PQ turns about P towards a limiting position. The straight line which has the limiting position of the secant is called the *tangent to the curve at the point P*. As long as there is a *limiting position for* the secant, it makes no difference whether the secant ever reaches that position or not, as the line having the limiting position is the tangent.

8. Slope of the Tangent.—In a smooth, continuous curve such as shown in Fig. 67, there must be one such limiting position at P and therefore, one tangent, such as PT. This tangent has a definite slope. Since, as the secant PQ turns on P, its varying slope always equals the *varying ratio* $\Delta y/\Delta x$, the limit value of $\Delta y/\Delta x$, as Q approaches P as a limiting position, will be the slope of the tangent at P.

9. There are points on certain curves such that the limit value of $\Delta y/\Delta x$ is not unique or may not exist. Such curves will not be taken up in this text. The further developments in this appendix are understood to apply to smooth continuous curves that have no such peculiar points.

10. Function.—Suppose the equation of the curve is given as y equals some function of x, written

$$y = f(x).$$

This is a generalization which simply means that y equals some *mathematical expression involving* x, such as $y = 3x - 1$, $y = \frac{1}{4}x^2 + 2$, $y = 2 \log x$, $y = 1/x$, etc. If the x-coordinate of some point on the curve is given, the corresponding y-coordinate is, at once, found by putting the given value for x in the function and computing the resulting value of the function. For example, in $y = \frac{1}{4}x^2 + 2$, if $x = 2$, $y = \frac{1}{4}(2)^2 + 2 = 3$. Thus the point (2, 3) is on the curve represented by $y = \frac{1}{4}x^2 + 2$.

Suppose, then, that on the curve $y = f(x)$, a point P is taken whose coordinates are x and y. Now change x by an increment

Δx, y will also change by a corresponding increment Δy, giving another point Q on the curve whose coordinates are $x + \Delta x$ and $y + \Delta y$. The value of $y + \Delta y$ is found by putting $x + \Delta x$ for x in the function so that

$$y + \Delta y = f(x + \Delta x).$$

In the above example,

$$y + \Delta y = \tfrac{1}{4}(x + \Delta x)^2 + 2.$$

The ratio $\Delta y/\Delta x$ is the slope of the secant through P and Q.

11. Slope of the Tangent, Derivative.—*To find the slope of the tangent at P, one must be able to find the limit value of $\Delta y/\Delta x$ as Δx approaches zero,* that is, as Q moves along the curve toward P as a limiting position. This is expressed thus:

$$\lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \frac{dy}{dx}.$$

It is read: "The limit of $\Delta y/\Delta x$ as Δx approaches 0 as a limit is dy/dx." This quantity, dy/dx, is called the *derivative of y with respect to x.* It is read "derivative of y with respect to x" or simply "dy over dx." The process of finding the derivative is called *differentiation.* This limit value is, in general, a function of x, the abscissa of the point at which the tangent is drawn. So, to obtain the slope of the tangent at any point whose abscissa is known, substitute the abscissa of the point for x in dy/dx, the derived function. As with Δ in Δx, so here the d is not a quantity, but merely a symbol of the result of a process. The derivative, dy/dx, is of fundamental importance in differential calculus and its applications.

12. Derivative of x^n.—In the calculus, rules are derived for writing down the derivatives of different types of functions. As an example of one method of attack for obtaining these rules, take the function

$$y = x^2.$$

Change x by the increment Δx. The function becomes

$$y + \Delta y = (x + \Delta x)^2 = x^2 + 2x\Delta x + (\Delta x)^2$$

Subtracting

$$y = x^2$$

leaves

$$\Delta y = 2x\Delta x + (\Delta x)^2.$$

Then dividing by Δx,

$$\frac{\Delta y}{\Delta x} = 2x + \Delta x.$$

It follows:

$$\frac{dy}{dx} = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = 2x,$$

since as Δx approaches 0 as a limit, $2x + \Delta x$ will approach $2x$ as a limit.

Take the more general function

$$y = x^n.$$

$$y + \Delta y = (x + \Delta x)^n.$$

$$y + \Delta y = x^n + nx^{n-1}\Delta x + \frac{n(n-1)}{\underline{|2}} x^{n-2}(\Delta x)^2 + \textit{terms in higher powers of } \Delta x.$$

Subtracting $y = x^n$, leaves

$$\Delta y = nx^{n-1}\Delta x + \frac{n(n-1)}{\underline{|2}} x^{n-2}(\Delta x)^2 + \textit{terms in higher powers of } \Delta x.$$

Dividing by Δx,

$$\frac{\Delta y}{\Delta x} = nx^{n-1} + \frac{n(n-1)}{\underline{|2}} x^{n-2}\Delta x + \textit{terms in higher powers of } \Delta x.$$

Since Δx may be made as small as we please, Δx times any finite quantity may be made as small as we please. When Δx becomes less than unity, as it approaches 0 as a limit, higher powers of Δx will approach 0 more rapidly than does Δx itself. Then the terms following nx^{n-1} each approach 0 as a limit. If there is a finite number of these terms, or if their sum is absolutely convergent, the sum approaches 0 as a limit. Therefore the limit,

$$\frac{dy}{dx} = nx^{n-1}.$$

This result translated into English reads: "The derivative of the nth power of x is n times x with the exponent diminished by one." This gives a rule for differentiating powers of x which may be proved to hold when n is negative or fractional as well as integral.

If

$$y = ax^n,$$

each term of the above expansion of $(x + \Delta x)^n$ is multiplied by a and

$$\frac{dy}{dx} = anx^{n-1}.$$

The rules for differentiation of other functions together with ıeir derivation are given in works on differential calculus.

The rule given above makes the derivative of a constant equal) 0. The constant a may be written ax^0. Now applying the ıle, the derivative is equal to $a \cdot 0 \cdot x^{-1}$, which equals 0. Or, $y = a$, a constant, any change in x does not affect y and so Δy ould be 0 and $\Delta y/\Delta x = 0$. This remains 0 all the time as Δx pproaches 0 as a limit. Hence $dy/dx = 0$ when y equals a onstant.

13. Application.—Take again the function

$$y = \tfrac{1}{4}x^2 + 2.$$

Applying the rule for differentiation,

$$\frac{dy}{dx} = \frac{2}{4}x = \frac{x}{2}.$$

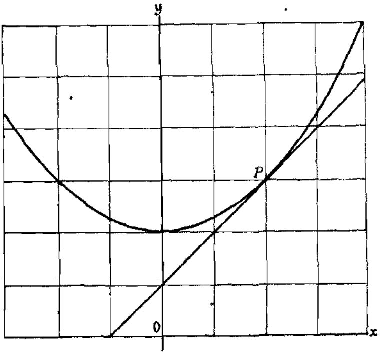


FIG. 68.—Tangent to $y = \tfrac{1}{4}x^2 + 2$.

Since the slope of the tangent to the curve, which is the graph of $y = f(x)$, at any point on the curve whose x-coordinate is given, is the value of dy/dx with the given value of x substituted for x in the derivative, it is seen at once that the slope of the tangent to the curve

$$y = \tfrac{1}{4}x^2 + 2$$

at the point whose x-coordinate is x_1 is $x_1/2$.

From the tabulation of values of y for varying values of x, the curve of Fig. 68 is plotted.

x	$y = \frac{1}{4}x^2 + 2$
-3	$4\frac{1}{4}$
-2	3
-1	$2\frac{1}{4}$
0	2
1	$2\frac{1}{4}$
2	3
3	$4\frac{1}{4}$
4	6

At $x = 2$ for the point P, $dy/dx = 2/2 = 1$. Now the tangent may be drawn at P with a slope of 1. A line with slope 1 rises vertically the same amount that it proceeds horizontally to the right.

14. Maximum.—Recall that the y-coordinate known as ordinate, of a point on a curve determined by $y = f(x)$ is the value of

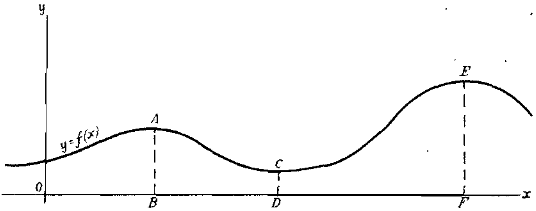


FIG. 69.—Maximum and minimum.

the function for that value of x which is the x-coordinate, or abscissa, of that point. A function of x is said to have a *maximum* value for that value x_1, of x, which makes the function larger than for other values of x in the neighborhood of x_1 either greater or less than x_1.

In Fig. 69 at $x = x_1 = OB$, the value of the function represented by the curve is $y = BA$. This is greater than the ordinate of a point on the curve in the neighborhood of A on either side of A. Then $y = BA$ is a maximum for $x = x_1 = OB$. The function is said to be a *maximum* for $x = OB$. Similarly, at $x = OF$, there is another point E on the curve for which the same statement is

true; $y = FE$ is then a maximum value of the function. This curve has two maxima. This merely means that there are two points on the curve, each of which is higher than any points of the curve in its immediate neighborhood. It is said that $x = OB$ makes $f(x)$ a maximum and $x = OF$ makes $f(x)$ a maximum. The maximum values of the function are, of course, found by substituting for x in the function the value of x that makes the function a maximum.

15. Minimum.—In like manner, $x = OD$ determines a point C on the curve such that ordinates in its immediate neighborhood, on both sides of C, are greater than DC, the ordinate at C. Then $y = DC$ is called a *minimum* value of the function. The function is said to be a minimum for $x = OD$. The minimum value of the function is found by substituting OD for x in the function.

Some functions of x have neither a maximum nor a minimum for any value of x. Some may have several maxima and several minima.

16. With the sort of functions considered in this appendix, the tangent to the graph at maximum or minimum points is horizontal, or parallel to the x-axis. In other words, its slope is 0. At a maximum or a minimum point, then, there is a value of x which makes the derivative of the function zero.

17. Increasing and Decreasing Functions.—At any point on the graph of $y = f(x)$ where the curve is going up to the right,

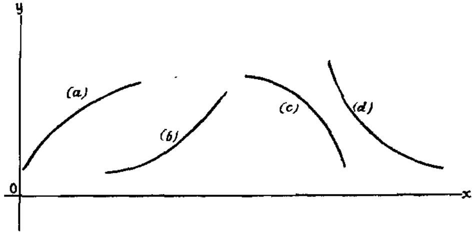


FIG. 70.—Increasing and decreasing functions.

the function is said to be *increasing* for the corresponding value of x, since it increases with increase of x. At any point where the curve is going down to the right, the function is said to be a *decreasing* function for the corresponding value of x.

In Fig. 70, (*a*) and (*b*) represent increasing functions for all values of x shown, while (*c*) and (*d*) represent decreasing functions. Note that at any point on either (*a*) or (*b*) the tangent slopes up to the right, that is, has a positive slope. This means that if the x-coordinate of any point on such a curve be substituted for x in the derivative of the function, the result will be a positive quantity. Similarly, for a decreasing function at any value of x, the derivative is negative.

18. Test for Maximum or Minimum.—If, on the curve $y = f(x)$ in Fig. 69 we pass tothe right approaching the high point A, the slope of the tangent is positive, decreasing to zero at A; and then continuing to decrease, it becomes negative. Since the derivative of $f(x)$ determines the slope of the tangent, it now follows that as the high point A is approached and passed through, the derivative is a decreasing function, passing from positive to negative through the value 0 at the high point A. It follows then that if the x-coordinate of A, a high point, be substituted for x in the derivative of the derivative, a negative result will be obtained. The derivative of the derivative of $y = f(x)$ is called the *second derivative* of the function. A symbol for the second derivative of y with respect to x is d^2y/dx^2; dy/dx is called the *first derivative*.

A similar argument will show that if the x-coordinate of the low point C be substituted for x in d^2y/dx^2, the result will be positive, since at C the slope of the tangent, dy/dx, is an increasing function, changing from negative to zero at C and then becoming positive.

19. Rules.—Rules may now be stated for determining maximum and minimum values of the kind of functions being discussed.

Having given $y = f(x)$, find dy/dx. Set $dy/dx = 0$ and solve for x. The resulting value or values of x determine the high points and low points on the graph of $y = f(x)$, since at those points the slope of the tangent is zero. If x_1 is one such value of x, substitute x_1 for x in $f(x)$. The result is the corresponding maximum or minimum value of the function as the case may be.

To determine whether the resulting value is a maximum or a minimum, find d^2y/dx^2. Substitute x_1 for x in d^2y/dx^2. If the result is negative, a maximum value of $f(x)$ was determined. If the result is positive, a minimum value of $f(x)$ was determined. If the result is zero, neither positive nor negative, further investigation is necessary.

20. Illustrative Example.—Take, for example, the function

$$y = x^3 - 6x^2 + 9x + 2.$$

Differentiating gives

$$\frac{dy}{dx} = 3x^2 - 12x + 9.$$

Set

$$3x^2 - 12x + 9 = 0.$$

Solving for x gives

$$x = 1 \text{ or } 3.$$

A second differentiation gives

$$\frac{d^2y}{dx^2} = 6x - 12.$$

For $x = 1$,

$$\frac{d^2y}{dx^2} = 6 \cdot 1 - 12,$$

is negative. Therefore, y has a maximum value at $x = 1$. This value is

$$y = 1^3 - 6 \cdot 1^2 + 9 \cdot 1 + 2 = 6.$$

For $x = 3$,

$$\frac{d^2y}{dx^2} = 6 \cdot 3 - 12,$$

is positive. Therefore, y has a minimum value at $x = 3$. This value is

$$y = 3^3 - 6 \cdot 3^2 + 9 \cdot 3 + 2 = 2\cdot$$

The graph of the curve is shown in Fig. 71. This shows the maximum value of 6 at $x = 1$ and the minimum of 2 at $x = 3$.

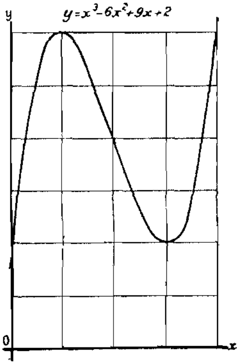


Fig. 71.

21. Speed and Acceleration.—The first derivative is used to determine the rate of change of a function. If x represents time and y space passed over in time x, represent the variables by t and s. An equational relationship between space passed over in a given time takes the general form

$$s = f(t).$$

Then Δs is the space passed over in time Δt, and $\Delta s/\Delta t$ is the average speed or velocity during time Δt.

$$\frac{ds}{dt} = \lim_{\Delta t \to 0} \frac{\Delta s}{\Delta t}$$

is defined as the instantaneous speed or velocity at time t. That is, ds/dt is the time rate of change of s at any instant t. If v stands for velocity, then

$$v = \frac{ds}{dt}.$$

Similarly, acceleration is the rate of change of velocity. Then acceleration is

$$a = \frac{dv}{dt} = \frac{d^2s}{dt^2}.$$

A body falling under the law

$$s = \tfrac{1}{2}gt^2$$

has at any instant a velocity

$$v = \frac{ds}{dt} = gt,$$

and an acceleration

$$a = \frac{dv}{dt} = \frac{d^2s}{dt^2} = g.$$

The constant g is called the gravitational constant.

22. A Function of a Function.—If y is a function of z and z is a function of x, dy/dx may be found directly. If x be given an increment, Δx, z takes an increment Δz. Then y takes an increment Δy. Now

$$\frac{\Delta y}{\Delta x} = \frac{\Delta y}{\Delta z} \cdot \frac{\Delta z}{\Delta x}.$$

It can be shown that the limit of the product of two variables is the product of their limits. Therefore, from the above equation.

$$\frac{dy}{dx} = \frac{dy}{dz} \cdot \frac{dz}{dx}.$$

As an illustrative example, let $y = z^3$ and $z = x^2 - 2x + 1$. Then,

$$\frac{dy}{dz} = 3z^2,$$

$$\frac{dz}{dx} = 2x - 2,$$

and

$$\frac{dy}{dx} = \frac{dy}{dz} \cdot \frac{dz}{dx}$$
$$= 3z^2(2x - 2)$$
$$= 3(x^2 - 2x + 1)^2(2x - 2).$$

23. This rule for differentiating a function of a function could be applied directly to such a function as

$$y = (x^2 - 2x + 1)^3.$$

Let

$$z = (x^2 - 2x + 1).$$

Then,

$$y = z^3,$$

and exactly the above example results. Write the following rule for differentiation of a function of x of the form y equals a power of a function of x.

Multiply the function of x by the index of the power, diminish the index of the power by 1, and multiply the result by the derivative of the function of x.

As an example, take

$$y = (x^3 - 3x^2 + 2x - 1)^4.$$
$$\frac{dy}{dx} = 4(x^3 - 3x^2 + 2x - 1)^3(3x^2 - 6x + 2).$$

24. A Function of Two Variables.—Sometimes y may be given as a function of two variables x and z. In such a case, to get a *geometric interpretation, three coordinate planes may be* used, mutually perpendicular. These planes intersect in three mutually perpendicular lines which are used as coordinate axes. Let Ox (Fig. 72) be the x-axis and Oz be the z-axis, both in the horizontal plane, and let Oy be the y-axis which is vertical. The coordinate planes are the horizontal plane, xOz, the vertical plane, xOy, and the vertical plane, zOy. Suppose the latter to be perpendicular to the plane of the paper.

25. Three Coordinates.—A point in space has three coordinates, x, y, and z, written (x, y, z). For example, the point $(1, 2, 3)$ is one unit distant from the zOy-plane, two units distant from the xOz-plane, and three units distant from the xOy-plane.

Let the positive direction for x be to the right, for y upward, and for z toward the front. The opposite directions are then negative.

26. Graph of $f(x, z)$.—An equation giving y as a function of x and z,

$$y = f(x, z),$$

represents a surface. As soon as values are assigned to x and z, y is determined. In Fig. 72, for $x = OD = x_1$ and $z = OH = z_1$ the point E is determined in the xOz-plane. If $y = f(x, z)$ is the equation of the surface ABC, then when $OD = x_1$ is substituted for x, and $OH = z_1$ for z in the function, the value $y = EP$

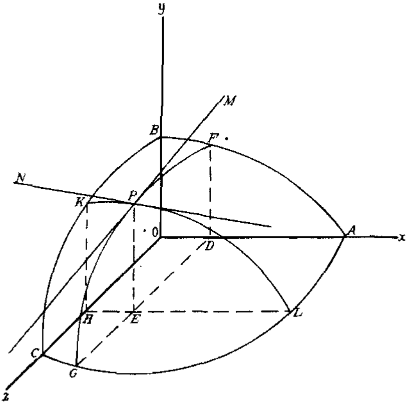


FIG. 72.—Three-dimensional coordinates.

$= y_1$ is at once known, determining the point P on the surface. This is expressed by

$$y_1 = f(x_1, z_1).$$

27. Partial Derivative.—Every point in the plane FDG, passed perpendicular to Ox at D, has an x-coordinate equal to $OD = x_1$. This plane cuts the surface in the curve FPG. The equation of this curve in the plane FDG is

$$y = f(x_1, z),$$

in which x_1 is a constant equal to OD. The slope of the tangent to this curve in the plane FDG at a point on the curve is the derivative of y with respect to z, letting x remain constant and equal

to x_1. Such a derivative is called partial derivative of y with respect to z. Its symbol is $\delta y/\delta z$. Similarly, the slope of the tangent to the curve KPL, in the plane KHL, at a point on the curve is $\delta y/\delta x$, letting z remain constant and equal to $OH = z_1$.

28. Conditions Necessary for Maximum or Minimum.—Now, in $\delta y/\delta z$, substitute for z the z-coordinate of the point P, and the result gives the slope of the tangent to the curve FPG at P. Let this tangent be PM. Similarly, in $\delta y/\delta x$, substitute the x-coordinate of P for x, and the result gives the slope of the tangent to the curve KPL at P. Let this tangent be PN.

If $EP = y_1$ were a maximum value of the function $f(x, y)$, it is evident that it would have to be a maximum for the curve $y = f(x_1, z)$ and the curve $y = f(x, z_1)$. This means that the tangents PM and PN would both be horizontal. In other words, necessary conditions for a maximum would be

$$\frac{\delta y}{\delta x} = 0,$$

$$\frac{\delta y}{\delta z} = 0.$$

Evidently these conditions must also hold for a minimum.

The partial derivative $\delta y/\delta x$ may be again differentiated with respect to x leaving z constant, or with respect to z leaving x constant. If the differentiations are both with respect to x leaving z constant, the symbol is $\delta^2 y/\delta x^2$. If the first differentiation is with respect to x and the second one with respect to z, the symbol is $\delta^2 y/\delta x\ \delta z$.

For either a maximum or a minimum, it can be shown that $(\delta^2 y/\delta x \delta z)^2$ must be less than $\dfrac{\delta^2 y}{\delta x^2} \cdot \dfrac{\delta^2 y}{\delta z^2}$.

So the conditions for either a maximum or a minimum are

$$\frac{\delta y}{\delta x} = 0,$$

$$\frac{\delta y}{\delta z} = 0,$$

$$\left(\frac{\delta^2 y}{\delta x \delta z}\right)^2 < \frac{\delta^2 y}{\delta x^2} \cdot \frac{\delta^2 y}{\delta z^2}.$$

In addition to these conditions, add that for a maximum

$$\frac{\delta^2 y}{\delta x^2} < 0 \text{ and } \frac{\delta^2 y}{\delta z^2} < 0,$$

and for a minimum

$$\frac{\delta^2 y}{\delta x^2} > 0 \text{ and } \frac{\delta^2 y}{\delta z^2} > 0,$$

at the point in question. These conditions may be taken without further proof.

29. Illustrative Example.—As an example, test

$$y = 3xz - x^3 - z^3$$

for maxima and minima.

$$\frac{\delta y}{\delta x} = 3z - 3x^2$$

$$\frac{\delta y}{\delta z} = 3x - 3z^2.$$

Equating each to zero, gives

$$z - x^2 = 0,$$
$$x - z^2 = 0.$$

Solving for x and z gives

$$x = 0 \text{ or } 1,$$
$$z = 0 \text{ or } 1.$$

Now,

$$\frac{\delta^2 y}{\delta x^2} = -6x,$$

$$\frac{\delta^2 y}{\delta z^2} = -6z,$$

$$\frac{\delta^2 y}{\delta x \delta z} = 3.$$

For $x = 0$ and $z = 0$,

$$\frac{\delta^2 y}{\delta x^2} = 0, \frac{\delta^2 y}{\delta z^2} = 0, \frac{\delta^2 y}{\delta x \delta z} = 3.$$

$$\left(\frac{\delta^2 y}{\delta x \delta z}\right)^2 = 9$$

is not less than

$$\frac{\delta^2 y}{\delta x^2} \cdot \frac{\delta^2 y}{\delta z^2} = 0.$$

Therefore, there is neither a maximum nor minimum at $x = 0$, $z = 0$.

For $x = 1$ and $z = 1$,

$$\frac{\delta^2 y}{\delta x^2} = -6, \frac{\delta^2 y}{\delta z^2} = -6, \frac{\delta^2 y}{\delta x \delta z} = 3.$$

$$\left(\frac{\delta^2 y}{\delta x \delta z}\right)^2 = 9, \frac{\delta^2 y}{\delta x^2} \cdot \frac{\delta^2 y}{\delta z^2} = 36.$$

Thus the conditions for a maximum or a minimum are fulfilled at $x = 1$, $z = 1$. Both $\delta^2 y/\delta x^2$ and $\delta^2 y/\delta z^2$ at this point are negative. Therefore, there is a maximum value of y at $x = 1$, $z = 1$. This maximum value is obtained by substituting 1 for x and 1 for z in the original equation. This gives

$$y = 3 \cdot 1 \cdot 1 - 1^3 - 1^3 = 1$$

as the maximum value. This means that if the surface represented by the original equation were constructed, the point on the surface at $x = 1$, $z = 1$ would be higher than any points in its immediate neighborhood. It would be at a height of 1.

30. Application to Least Squares Adjustment.—In Chap. VIII, Sec. 8, *the method of finding the straight line of closest fit* by least squares adjustment was found to depend on finding the values of a and b such that

$$\Sigma(ax + b - y)^2$$

should be a minimum. Now find the required values of a and b by the above conditions and tests.

Let

$$u = \Sigma(ax + b - y)^2.$$

The unknown quantities to be determined are the values of a and b to make u a minimum.

$$\frac{\delta u}{\delta b} = \Sigma 2(ax + b - y),$$

$$\frac{\delta u}{\delta a} = \Sigma 2(ax + b - y)x.$$

The equations to solve for a and b are then

$$\Sigma 2(ax + b - y) = 0$$

and

$$\Sigma 2(ax + b - y)x = 0.$$

These reduce to

$$a\Sigma x + Nb - \Sigma y = 0,$$
$$a\Sigma x^2 + b\Sigma x - \Sigma xy = 0.$$

Solving for a and b gives

$$a = \frac{\Sigma x \cdot \Sigma y - N\Sigma xy}{(\Sigma x)^2 - N\Sigma x^2},$$

$$b = \frac{\Sigma x \cdot \Sigma xy - \Sigma x^2 \cdot \Sigma y}{(\Sigma x)^2 - N\Sigma x^2}.$$

Now

$$\frac{\delta^2 u}{\delta b \delta a} = 2\Sigma x.$$

$$\frac{\delta^2 u}{\delta b^2} = 2N.$$

$$\frac{\delta^2 u}{\delta a^2} = 2\Sigma x^2.$$

The condition

$$\left(\frac{\delta^2 u}{\delta b \delta a}\right)^2 < \frac{\delta^2 u}{\delta b^2} \cdot \frac{\delta^2 u}{\delta a^2},$$

gives

$$(2\Sigma x)^2 < 2N \cdot 2\Sigma x^2.$$

It is easily shown that if x is increasing by constant differences,

$$(2\Sigma x)^2 < 2N \cdot 2\Sigma x^2.$$

Moreover, $\delta^2 u/\delta b^2 = 2N$ and $\delta^2 u/\delta a^2 = 2\Sigma x^2$ are each positive, regardless of the values of a and b. It thus appears that the conditions are all fulfilled for giving a minimum value to u at the values of a and b obtained above.

This completes a brief consideration of two important applications of differentiation, namely: (1) determination of maximum and minimum values, and (2) rates of change.

31. Integration, Constant of Integration.—Integration is the inverse process of differentiation, just as subtraction is the inverse of addition, division the inverse of multiplication. In other words, to *integrate* $y = f(x)$, find a function which differentiated will give $f(x)$.

For example, the *integral* of $y = x^2$ is $x^3/3$, since $x^3/3$ differentiated gives $3x^2/3 = x^2$. But $x^3/3 + 2$, $x^3/3 - 6$, and $x^3/3$ plus any constant C also differentiate into x^2. For this reason, when integrating a function, add a constant C to the result. This is called the *constant of integration.* Unless there are conditions which determine this constant, it may be any constant whatever.

The symbol of integration is an elongated *s*. The integral of (x) is written

$$\int f(x)dx.$$

Thus

$$\int x^2 dx = \frac{x^3}{3} + C.$$

This is read "the integral of x square dx equals $x^3/3$ plus a constant."

In general,

$$\int x^n dx = \frac{x^{n+1}}{n+1} + C.$$

This holds except for $n = -1$.

32. Just as a constant factor in a function to be differentiated appears in the derivative, so a constant factor in a function to be integrated appears in the resulting integral. Since the constant of integration may be any constant, the constant factor may be regarded as a factor of it also. It follows that a constant factor may be placed either side of the integral sign without changing the result.

33. Geometric Meaning.—To get a geometric interpretation of the meaning of integration, suppose $y = f(x)$ to be the equation

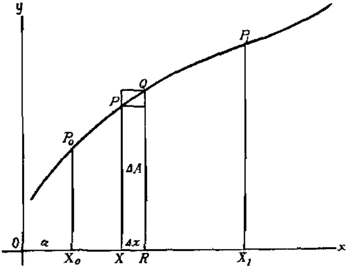


Fig. 73.—Geometric meaning of integration.

of the curve in Fig. 73. If the point P starts at P_0, where $X = OX_0 = a$, and moves according to the law $y = f(x)$ till $x = OX_1 = b$, P may be said to generate the curve from P_0 to P_1. At the same time, the ordinate of P, XP, sweeps over or generates the area $P_0X_0X_1P_1$ under the curve.

In taking any position of the generating point such as P, the ordinate XP at that point is $y = f(x)$, where $x = OX$. The area generated up to XP is P_0X_0XP. Now give x an increment $\Delta x = XR$. The area A takes an increment ΔA. This increment of area is $PXRQ$ between the curve, the x-axis, and the ordinates

XP and RQ. The area ΔA as shown is greater than the rectangle PR and less than the rectangle QX. Always

$$\Delta A > \Delta x \text{ times the least ordinate to arc } PQ$$

and

$$\Delta A < \Delta x \text{ times the greatest ordinate to arc } PQ$$

So that

$$\frac{\Delta A}{\Delta x} > \text{the least ordinate to arc } PQ,$$

and

$$\frac{\Delta A}{\Delta x} < \text{the greatest ordinate to arc } PQ.$$

In the figure, the least ordinate is XP and the greatest is RQ. In any case as $\Delta x \to 0$, the difference between the least ordinate and $y = XP$ as well as the difference between the greatest ordinate and $y = XP$ may be made to become and remain less than any assignable quantity, no matter how small. In other words, if $\Delta x \to 0$, the limit of each ordinate in question is $y = XP$. Since, as $\Delta x \to 0$, the value of $\Delta A/\Delta x$ remains between the values of the greatest and least ordinates to arc PQ, the limit of $\Delta A/\Delta x$ must be $y = XP$, the common limit of the two ordinates.

Therefore,

$$\lim_{\Delta x \to 0} \frac{\Delta A}{\Delta x} = y,$$

or

$$\frac{dA}{dx} = y.$$

Hence, by integration, area

$$A = \int y dx = \int f(x) dx.$$

Let

$$\int f(x) dx = F(x) + C.$$

Then

$$A = F(x) + C.$$

The value of the constant of integration, C, is determined by the initial conditions at $x = OX_0 = a$. When $x = a$, the point generating the curve is at P_0 and the ordinate X_0P_0 has swept over no area at all. In other words, when $x = a$, $A = 0$.

Therefore,

$$0 = F(a) + C,$$

and

$$C = -F(a).$$

Substituting this value of C gives

$$A = F(x) - F(a)$$

as the area generated by the ordinate when P moves from P_0 to any point P.

Let $OX_1 = b$. Then the area $P_0X_0X_1P_1$ is

$$A = F(b) - F(a).$$

34. Definite Integral.—Starting, then, with the curve $y = f(x)$, the area between the curve, the x-axis, and the ordinates at $x = a$ and $x = b$ is expressed as

$$A = \int_a^b f(x)dx = F(b) - F(a),$$

in which $F(x)$ is the result of the integration $\int f(x)dx$.

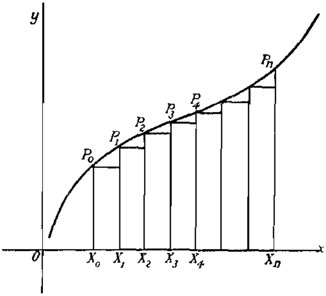


Fig. 74.

$\int_a^b f(x)dx$ is called the *definite integral* of $f(x)$ between the limits a and b.

It can be shown[1] that if the line X_0X_n be divided into segments of length Δx, and ordinates drawn at the points of division, and rectangles be completed by drawing horizontal lines from the top of each ordinate to the next ordinate as shown in Fig. 74, then as $\Delta x \rightarrow 0$ the sum of the areas of the rectangles approaches as a limit the area between the curve, the x-axis, and the ordinates at X_0 and X_n. *It follows that the limit of the sum of the areas of the rectangles, as the width of each approaches zero as a limit and their number increases without limit, is equal to the*

[1] See works on Integral Calculus.

definite integral $\int_a^b f(x)dx$, where $a = OX_0$, $b = OX_n$, and $y = f(x)$ is the equation of the curve.

In this manner, the definite integral is regarded as a limit of a sum and is useful in many problems of summation.

35. Illustrative Example.—As an example in integration, find the area under the curve in Fig. 71 from $x = 1$ to $x = 3$. The equation of the curve is

$$y = x^3 - 6x^2 + 9x + 2.$$

The area is

$$A = \int_1^3 (x^3 - 6x^2 + 9x + 2)dx$$

$$= \left[\frac{x^4}{4} - \frac{6x^3}{3} + \frac{9x^2}{2} + 2x + C\right]_1^3$$

$$= \left[\frac{3^4}{4} - \frac{6 \cdot 3^3}{3} + \frac{9 \cdot 3^2}{2} + 2 \cdot 3 + C\right] - \left[\frac{1}{4} - \frac{6}{3} + \frac{9}{2} + 2 + C\right]$$

$$= 8.$$

The limits of integration, 1 and 3, written with the bracket indicate that 1 is to be substituted for x within the bracket and the result subtracted from the result of substituting 3 for x within the bracket. Notice that C, the constant of integration, will always cancel out in the subtraction. For this reason C may be neglected in connection with the definite integral.

SQUARES OF NUMBERS

N	0	1	2	3	4	5	6	7	8	9	Avg. diff.
1.00	1.000	1.002	1.004	1.006	1.008	1.010	1.012	1.014	1.016	1.018	2
1	1.020	1.022	1.024	1.026	1.028	1.030	1.032	1.034	1.036	1.038	
2	1.040	1.042	1.044	1.047	1.049	1.051	1.053	1.055	1.057	1.059	
3	1.061	1.063	1.065	1.067	1.069	1.071	1.073	1.075	1.077	1.080	
4	1.082	1.084	1.086	1.088	1.090	1.092	1.094	1.096	1.098	1.100	
1.05	1.102	1.105	1.107	1.109	1.111	1.113	1.115	1.117	1.119	1.121	
6	1.124	1.126	1.128	1.130	1.132	1.134	1.136	1.138	1.141	1.143	
7	1.145	1.147	1.149	1.151	1.153	1.156	1.158	1.160	1.162	1.164	
8	1.166	1.169	1.171	1.173	1.175	1.177	1.179	1.182	1.184	1.186	
9	1.188	1.190	1.192	1.195	1.197	1.199	1.201	1.203	1.206	1.208	
1.10	1.210	1.212	1.214	1.217	1.219	1.221	1.223	1.225	1.228	1.230	
1	1.232	1.234	1.237	1.239	1.241	1.243	1.245	1.248	1.250	1.252	
2	1.254	1.257	1.259	1.261	1.263	1.266	1.268	1.270	1.272	1.275	
3	1.277	1.279	1.281	1.284	1.286	1.288	1.290	1.293	1.295	1.297	
4	1.300	1.302	1.304	1.306	1.309	1.311	1.313	1.316	1.318	1.320	
1.15	1.322	1.325	1.327	1.329	1.332	1.334	1.336	1.339	1.341	1.343	
6	1.346	1.348	1.350	1.353	1.355	1.357	1.360	1.362	1.364	1.367	
7	1.369	1.371	1.374	1.376	1.378	1.381	1.383	1.385	1.388	1.390	
8	1.392	1.395	1.397	1.399	1.402	1.404	1.407	1.409	1.411	1.414	
9	1.416	1.418	1.421	1.423	1.426	1.428	1.430	1.433	1.435	1.438	
1.20	1.440	1.442	1.445	1.447	1.450	1.452	1.454	1.457	1.459	1.462	
1	1.464	1.467	1.469	1.471	1.474	1.476	1.479	1.481	1.484	1.486	
2	1.488	1.491	1.493	1.496	1.498	1.501	1.503	1.506	1.508	1.510	
3	1.513	1.515	1.518	1.520	1.523	1.525	1.528	1.530	1.533	1.535	
4	1.538	1.540	1.543	1.545	1.548	1.550	1.553	1.555	1.558	1.560	
1.25	1.562	1.565	1.568	1.570	1.573	1.575	1.578	1.580	1.583	1.585	3
6	1.588	1.590	1.593	1.595	1.598	1.600	1.603	1.605	1.608	1.610	
7	1.613	1.615	1.618	1.621	1.623	1.626	1.628	1.631	1.633	1.636	
8	1.638	1.641	1.644	1.646	1.649	1.651	1.654	1.656	1.659	1.662	
9	1.664	1.667	1.669	1.672	1.674	1.677	1.680	1.682	1.685	1.687	
1.30	1.690	1.693	1.695	1.698	1.700	1.703	1.706	1.708	1.711	1.713	
1	1.716	1.719	1.721	1.724	1.727	1.729	1.732	1.734	1.737	1.740	
2	1.742	1.745	1.748	1.750	1.753	1.756	1.758	1.761	1.764	1.766	
3	1.769	1.772	1.774	1.777	1.780	1.782	1.785	1.788	1.790	1.793	
4	1.796	1.798	1.801	1.804	1.806	1.809	1.812	1.814	1.817	1.820	
1.35	1.822	1.825	1.828	1.831	1.833	1.836	1.839	1.841	1.844	1.847	
6	1.850	1.852	1.855	1.858	1.860	1.863	1.866	1.869	1.871	1.874	
7	1.877	1.880	1.882	1.885	1.888	1.891	1.893	1.896	1.899	1.902	
8	1.904	1.907	1.910	1.913	1.915	1.918	1.921	1.924	1.927	1.929	
9	1.932	1.935	1.938	1.940	1.943	1.946	1.949	1.952	1.954	1.957	
1.40	1.960	1.963	1.966	1.968	1.971	1.974	1.977	1.980	1.982	1.985	
1	1.988	1.991	1.994	1.997	1.999	2.002	2.005	2.008	2.011	2.014	
2	2.016	2.019	2.022	2.025	2.028	2.031	2.033	2.036	2.039	2.042	
3	2.045	2.048	2.051	2.053	2.056	2.059	2.062	2.065	2.068	2.071	
4	2.074	2.076	2.079	2.082	2.085	2.088	2.091	2.094	2.097	2.100	
1.45	2.102	2.105	2.108	2.111	2.114	2.117	2.120	2.123	2.126	2.129	
6	2.132	2.135	2.137	2.140	2.143	2.146	2.149	2.152	2.155	2.158	
7	2.161	2.164	2.167	2.170	2.173	2.176	2.179	2.182	2.184	2.187	
8	2.190	2.193	2.196	2.199	2.202	2.205	2.208	2.211	2.214	2.217	
9	2.220	2.223	2.226	2.229	2.232	2.235	2.238	2.241	2.244	2.247	

Moving the decimal point ONE place in *N* requires moving it TWO places in body of table.

SQUARES (*continued*)

N	0	1	2	3	4	5	6	7	8	9	Avg. diff.
1.50	2.250	2.253	2.256	2.259	2.262	2.265	2.268	2.271	2.274	2.277	3
1	2.280	2.283	2.286	2.289	2.292	2.295	2.298	2.301	2.304	2.307	
2	2.310	2.313	2.316	2.320	2.323	2.326	2.329	2.332	2.335	2.338	
3	2.341	2.344	2.347	2.350	2.353	2.356	2.359	2.362	2.365	2.369	
4	2.372	2.375	2.378	2.381	2.384	2.387	2.390	2.393	2.396	2.399	
1.55	2.402	2.406	2.409	2.412	2.415	2.418	2.421	2.424	2.427	2.430	
6	2.434	2.437	2.440	2.443	2.446	2.449	2.452	2.455	2.459	2.462	
7	2.465	2.468	2.471	2.474	2.477	2.481	2.484	2.487	2.490	2.493	
8	2.496	2.500	2.503	2.506	2.509	2.512	2.515	2.519	2.522	2.525	
9	2.528	2.531	2.534	2.538	2.541	2.544	2.547	2.550	2.554	2.557	
1.60	2.560	2.563	2.566	2.570	2.573	2.576	2.579	2.582	2.586	2.589	
1	2.592	2.595	2.599	2.602	2.605	2.608	2.611	2.615	2.618	2.621	
2	2.624	2.628	2.631	2.634	2.637	2.641	2.644	2.647	2.650	2.654	
3	2.657	2.660	2.663	2.667	2.670	2.673	2.676	2.680	2.683	2.686	
4	2.690	2.693	2.696	2.699	2.703	2.706	2.709	2.713	2.716	2.719	
1.65	2.722	2.726	2.729	2.732	2.736	2.739	2.742	2.746	2.749	2.752	
6	2.756	2.759	2.762	2.766	2.769	2.772	2.776	2.779	2.782	2.786	
7	2.789	2.792	2.796	2.799	2.802	2.806	2.809	2.812	2.816	2.819	
8	2.822	2.826	2.829	2.832	2.836	2.839	2.843	2.846	2.849	2.853	
9	2.856	2.859	2.863	2.866	2.870	2.873	2.876	2.880	2.883	2.887	
1.70	2.890	2.893	2.897	2.900	2.904	2.907	2.910	2.914	2.917	2.921	
1	2.924	2.928	2.931	2.934	2.938	2.941	2.945	2.948	2.952	2.955	
2	2.958	2.962	2.965	2.969	2.972	2.976	2.979	2.983	2.986	2.989	
3	2.993	2.996	3.000	3.003	3.007	3.010	3.014	3.017	3.021	3.024	
4	3.028	3.031	3.035	3.038	3.042	3.045	3.049	3.052	3.056	3.059	
1.75	3.062	3.066	3.070	3.073	3.077	3.080	3.084	3.087	3.091	3.094	4
6	3.098	3.101	3.105	3.108	3.112	3.115	3.119	3.122	3.126	3.129	
7	3.133	3.136	3.140	3.144	3.147	3.151	3.154	3.158	3.161	3.165	
8	3.168	3.172	3.176	3.179	3.183	3.186	3.190	3.193	3.197	3.201	
9	3.204	3.208	3.211	3.215	3.218	3.222	3.226	3.229	3.233	3.236	
1.80	3.240	3.244	3.247	3.251	3.254	3.258	3.262	3.265	3.269	3.272	
1	3.276	3.280	3.283	3.287	3.291	3.294	3.298	3.301	3.305	3.309	
2	3.312	3.316	3.320	3.323	3.327	3.331	3.334	3.338	3.342	3.345	
3	3.349	3.353	3.356	3.360	3.364	3.367	3.371	3.375	3.378	3.382	
4	3.386	3.389	3.393	3.397	3.400	3.404	3.408	3.411	3.415	3.419	
1.85	3.422	3.426	3.430	3.434	3.437	3.441	3.445	3.448	3.452	3.456	
6	3.460	3.463	3.467	3.471	3.474	3.478	3.482	3.486	3.489	3.493	
7	3.497	3.501	3.504	3.508	3.512	3.516	3.519	3.523	3.527	3.531	
8	3.534	3.538	3.542	3.546	3.549	3.553	3.557	3.561	3.565	3.568	
9	3.572	3.576	3.580	3.583	3.587	3.591	3.595	3.599	3.602	3.606	
1.90	3.610	3.614	3.618	3.621	3.625	3.629	3.633	3.637	3.640	3.644	
1	3.648	3.652	3.656	3.660	3.663	3.667	3.671	3.675	3.679	3.683	
2	3.686	3.690	3.694	3.698	3.702	3.706	3.709	3.713	3.717	3.721	
3	3.725	3.729	3.733	3.736	3.740	3.744	3.748	3.752	3.756	3.760	
4	3.764	3.767	3.771	3.775	3.779	3.783	3.787	3.791	3.795	3.799	
1.95	3.802	3.806	3.810	3.814	3.818	3.822	3.826	3.830	3.834	3.838	
6	3.842	3.846	3.849	3.853	3.857	3.861	3.865	3.869	3.873	3.877	
7	3.881	3.885	3.889	3.893	3.897	3.901	3.905	3.909	3.912	3.916	
8	3.920	3.924	3.928	3.932	3.936	3.940	3.944	3.948	3.952	3.956	
9	3.960	3.964	3.968	3.972	3.976	3.980	3.984	3.988	3.992	3.996	

$\pi^2 = 9.86960$ $1/\pi^2 = 0.101321$ $e^2 = 7.38906$

SQUARES (*continued*)

N	0	1	2	3	4	5	6	7	8	9	Avg. diff.
2.00	4.000	4.004	4.008	4.012	4.016	4.020	4.024	4.028	4.032	4.036	4
1	4.040	4.044	4.048	4.052	4.056	4.060	4.064	4.068	4.072	4.076	
2	4.080	4.084	4.088	4.093	4.097	4.101	4.105	4.109	4.113	4.117	
3	4.121	4.125	4.129	4.133	4.137	4.141	4.145	4.149	4.153	4.158	
4	4.162	4.166	4.170	4.174	4.178	4.182	4.186	4.190	4.194	4.198	
2.05	4.202	4.207	4.211	4.215	4.219	4.223	4.227	4.231	4.235	4.239	
6	4.244	4.248	4.252	4.256	4.260	4.264	4.268	4.272	4.277	4.281	
7	4.285	4.289	4.293	4.297	4.301	4.306	4.310	4.314	4.318	4.322	
8	4.326	4.331	4.335	4.339	4.343	4.347	4.351	4.356	4.360	4.364	
9	4.368	4.372	4.376	4.381	4.385	4.389	4.393	4.397	4.402	4.406	
2.10	4.410	4.414	4.418	4.423	4.427	4.431	4.435	4.439	4.444	4.448	
1	4.452	4.456	4.461	4.465	4.469	4.473	4.477	4.482	4.486	4.490	
2	4.494	4.499	4.503	4.507	4.511	4.516	4.520	4.524	4.528	4.533	
3	4.537	4.541	4.545	4.550	4.554	4.558	4.562	4.567	4.571	4.575	
4	4.580	4.584	4.588	4.592	4.597	4.601	4.605	4.610	4.614	4.618	
2.15	4.622	4.627	4.631	4.635	4.640	4.644	4.648	4.653	4.657	4.661	
6	4.666	4.670	4.674	4.679	4.683	4.687	4.692	4.696	4.700	4.705	
7	4.709	4.713	4.718	4.722	4.726	4.731	4.735	4.739	4.744	4.748	
8	4.752	4.757	4.761	4.765	4.770	4.774	4.779	4.783	4.787	4.792	
9	4.796	4.800	4.805	4.809	4.814	4.818	4.822	4.827	4.831	4.836	
2.20	4.840	4.844	4.849	4.853	4.858	4.862	4.866	4.871	4.875	4.880	
1	4.884	4.889	4.893	4.897	4.902	4.906	4.911	4.915	4.920	4.924	
2	4.928	4.933	4.937	4.942	4.946	4.951	4.955	4.960	4.964	4.968	
3	4.973	4.977	4.982	4.986	4.991	4.995	5.000	5.004	5.009	5.013	
4	5.018	5.022	5.027	5.031	5.036	5.040	5.045	5.049	5.054	5.058	
2.25	5.062	5.067	5.072	5.076	5.081	5.085	5.090	5.094	5.099	5.103	5
6	5.108	5.112	5.117	5.121	5.126	5.130	5.135	5.139	5.144	5.148	
7	5.153	5.157	5.162	5.167	5.171	5.176	5.180	5.185	5.189	5.194	
8	5.198	5.203	5.208	5.212	5.217	5.221	5.226	5.230	5.235	5.240	
9	5.244	5.249	5.253	5.258	5.262	5.267	5.272	5.276	5.281	5.285	
2.30	5.290	5.295	5.299	5.304	5.308	5.313	5.318	5.322	5.327	5.331	
1	5.336	5.341	5.345	5.350	5.355	5.359	5.364	5.368	5.373	5.378	
2	5.382	5.387	5.392	5.396	5.401	5.406	5.410	5.415	5.420	5.424	
3	5.429	5.434	5.438	5.443	5.448	5.452	5.457	5.462	5.466	5.471	
4	5.476	5.480	5.485	5.490	5.494	5.499	5.504	5.508	5.513	5.518	
2.35	5.522	5.527	5.532	5.537	5.541	5.546	5.551	5.555	5.560	5.565	
6	5.570	5.574	5.579	5.584	5.588	5.593	5.598	5.603	5.607	5.612	
7	5.617	5.622	5.626	5.631	5.636	5.641	5.645	5.650	5.655	5.660	
8	5.664	5.669	5.674	5.679	5.683	5.688	5.693	5.698	5.703	5.707	
9	5.712	5.717	5.722	5.726	5.731	5.736	5.741	5.746	5.750	5.755	
2.40	5.760	5.765	5.770	5.774	5.779	5.784	5.789	5.794	5.798	5.803	
1	5.808	5.813	5.818	5.823	5.827	5.832	5.837	5.842	5.847	5.852	
2	5.856	5.861	5.866	5.871	5.876	5.881	5.885	5.890	5.895	5.900	
3	5.905	5.910	5.915	5.919	5.924	5.929	5.934	5.939	5.944	5.949	
4	5.954	5.958	5.963	5.968	5.973	5.978	5.983	5.988	5.993	5.998	
2.45	6.002	6.007	6.012	6.017	6.022	6.027	6.032	6.037	6.042	6.047	
6	6.052	6.057	6.061	6.066	6.071	6.076	6.081	6.086	6.091	6.096	
7	6.101	6.106	6.111	6.116	6.121	6.126	6.131	6.136	6.140	6.145	
8	6.150	6.155	6.160	6.165	6.170	6.175	6.180	6.185	6.190	6.195	
9	6.200	6.205	6.210	6.215	6.220	6.225	6.230	6.235	6.240	6.245	

Moving the decimal point ONE place in *N* requires moving it TWO places in body of table.

SQUARES (*continued*)

N	0	1	2	3	4	5	6	7	8	9	Avg. diff.
2.50	6.250	6.255	6.260	6.265	6.270	6.275	6.280	6.285	6.290	6.295	5
1	6.300	6.305	6.310	6.315	6.320	6.325	6.330	6.335	6.340	6.345	
2	6.350	6.355	6.360	6.366	6.371	6.376	6.381	6.386	6.391	6.396	
3	6.401	6.406	6.411	6.416	6.421	6.426	6.431	6.436	6.441	6.447	
4	6.452	6.457	6.462	6.467	6.472	6.477	6.482	6.487	6.492	6.497	
2.55	6.502	6.508	6.513	6.518	6.523	6.528	6.533	6.538	6.543	6.548	
6	6.554	6.559	6.564	6.569	6.574	6.579	6.584	6.589	6.595	6.600	
7	6.605	6.610	6.615	6.620	6.625	6.631	6.636	6.641	6.646	6.651	
8	6.656	6.662	6.667	6.672	6.677	6.682	6.687	6.693	6.698	6.703	
9	6.708	6.713	6.718	6.724	6.729	6.734	6.739	6.744	6.750	6.755	
2.60	6.760	6.765	6.770	6.776	6.781	6.786	6.791	6.796	6.802	6.807	
1	6.812	6.817	6.823	6.828	6.833	6.838	6.843	6.849	6.854	6.859	
2	6.864	6.870	6.875	6.880	6.885	6.891	6.896	6.901	6.906	6.912	
3	6.917	6.922	6.927	6.933	6.938	6.943	6.948	6.954	6.959	6.964	
4	6.970	6.975	6.980	6.985	6.991	6.996	7.001	7.007	7.012	7.017	
2.65	7.022	7.028	7.033	7.038	7.044	7.049	7.054	7.060	7.065	7.070	
6	7.076	7.081	7.086	7.092	7.097	7.102	7.108	7.113	7.118	7.124	
7	7.129	7.134	7.140	7.145	7.150	7.156	7.161	7.166	7.172	7.177	
8	7.182	7.188	7.193	7.198	7.204	7.209	7.215	7.220	7.225	7.231	
9	7.236	7.241	7.247	7.252	7.258	7.263	7.268	7.274	7.279	7.285	
2.70	7.290	7.295	7.301	7.306	7.312	7.317	7.322	7.328	7.333	7.339	
1	7.344	7.350	7.355	7.360	7.366	7.371	7.377	7.382	7.388	7.393	
2	7.398	7.404	7.409	7.415	7.420	7.426	7.431	7.437	7.442	7.447	
3	7.453	7.458	7.464	7.469	7.475	7.480	7.486	7.491	7.497	7.502	
4	7.508	7.513	7.519	7.524	7.530	7.535	7.541	7.546	7.552	7.557	
2.75	7.562	7.568	7.574	7.579	7.585	7.590	7.596	7.601	7.607	7.612	6
6	7.618	7.623	7.629	7.634	7.640	7.645	7.651	7.656	7.662	7.667	
7	7.673	7.678	7.684	7.690	7.695	7.701	7.706	7.712	7.717	7.723	
8	7.728	7.734	7.740	7.745	7.751	7.756	7.762	7.767	7.773	7.779	
9	7.784	7.790	7.795	7.801	7.806	7.812	7.818	7.823	7.829	7.834	
2.80	7.840	7.846	7.851	7.857	7.862	7.868	7.874	7.879	7.885	7.890	
1	7.896	7.902	7.907	7.913	7.919	7.924	7.930	7.935	7.941	7.947	
2	7.952	7.958	7.964	7.969	7.975	7.981	7.986	7.992	7.998	8.003	
3	8.009	8.015	8.020	8.026	8.032	8.037	8.043	8.049	8.054	8.060	
4	8.066	8.071	8.077	8.083	8.088	8.094	8.100	8.105	8.111	8.117	
2.85	8.122	8.128	8.134	8.140	8.145	8.151	8.157	8.162	8.168	8.174	
6	8.180	8.185	8.191	8.197	8.202	8.208	8.214	8.220	8.225	8.231	
7	8.237	8.243	8.248	8.254	8.260	8.266	8.271	8.277	8.283	8.289	
8	8.294	8.300	8.306	8.312	8.317	8.323	8.329	8.335	8.341	8.346	
9	8.352	8.358	8.364	8.369	8.375	8.381	8.387	8.393	8.398	8.404	
2.90	8.410	8.416	8.422	8.427	8.433	8.439	8.445	8.451	8.456	8.462	
1	8.468	8.474	8.480	8.486	8.491	8.497	8.503	8.509	8.515	8.521	
2	8.526	8.532	8.538	8.544	8.550	8.556	8.561	8.567	8.573	8.579	
3	8.585	8.591	8.597	8.602	8.608	8.614	8.620	8.626	8.632	8.638	
4	8.644	8.649	8.655	8.661	8.667	8.673	8.679	8.685	8.691	8.697	
2.95	8.702	8.708	8.714	8.720	8.726	8.732	8.738	8.744	8.750	8.756	
6	8.762	8.768	8.773	8.779	8.785	8.791	8.797	8.803	8.809	8.815	
7	8.821	8.827	8.833	8.839	8.845	8.851	8.857	8.863	8.868	8.874	
8	8.880	8.886	8.892	8.898	8.904	8.910	8.916	8.922	8.928	8.934	
9	8.940	8.946	8.952	8.958	8.964	8.970	8.976	8.982	8.988	8.994	

$\pi^2 = 9.86960$ $1/\pi^2 = 0.101321$ $e^2 = 7.38906$

SQUARES (*continued*)

N	0	1	2	3	4	5	6	7	8	9	Avg. diff.
3.00	9.000	9.006	9.012	9.018	9.024	9.030	9.036	9.042	9.048	9.054	6
1	9.060	9.066	9.072	9.078	9.084	9.090	9.096	9.102	9.108	9.114	
2	9.120	9.126	9.132	9.139	9.145	9.151	9.157	9.163	9.169	9.175	
3	9.181	9.187	9.193	9.199	9.205	9.211	9.217	9.223	9.229	9.236	
4	9.242	9.248	9.254	9.260	9.266	9.272	9.278	9.284	9.290	9.296	
3.05	9.302	9.309	9.315	9.321	9.327	9.333	9.339	9.345	9.351	9.357	
6	9.364	9.370	9.376	9.382	9.388	9.394	9.400	9.406	9.413	9.419	
7	9.425	9.431	9.437	9.443	9.449	9.456	9.462	9.468	9.474	9.480	
8	9.486	9.493	9.499	9.505	9.511	9.517	9.523	9.530	9.536	9.542	
9	9.548	9.554	9.560	9.567	9.573	9.579	9.585	9.591	9.598	9.604	
3.10	9.610	9.616	9.622	9.629	9.635	9.641	9.647	9.653	9.660	9.666	
1	9.672	9.678	9.685	9.691	9.697	9.703	9.709	9.716	9.722	9.728	
2	9.734	9.741	9.747	9.753	9.759	9.766	9.772	9.778	9.784	9.791	
3	9.797	9.803	9.809	9.816	9.822	9.828	9.834	9.841	9.847	9.853	
4	9.860	9.866	9.872	9.878	9.885	9.891	9.897	9.904	9.910	9.916	
3.15	9.922	9.929	9.935	9.941	9.948	9.954	9.960	9.967	9.973	9.979	
6	9.986	9.992	9.998	10.005							6
3.1							9.99	10.05	10.11	10.18	6
2	10.24	10.30	10.37	10.43	10.50	10.56	10.63	10.69	10.76	10.82	
3	10.89	10.96	11.02	11.09	11.16	11.22	11.29	11.36	11.42	11.49	7
4	11.56	11.63	11.70	11.76	11.83	11.90	11.97	12.04	12.11	12.18	
3.5	12.25	12.32	12.39	12.46	12.53	12.60	12.67	12.74	12.82	12.89	
6	12.96	13.03	13.10	13.18	13.25	13.32	13.40	13.47	13.54	13.62	
7	13.69	13.76	13.84	13.91	13.99	14.06	14.14	14.21	14.29	14.36	8
8	14.44	14.52	14.59	14.67	14.75	14.82	14.90	14.98	15.05	15.13	
9	15.21	15.29	15.37	15.44	15.52	15.60	15.68	15.76	15.84	15.92	
4.0	16.00	16.08	16.16	16.24	16.32	16.40	16.48	16.56	16.65	16.73	
1	16.81	16.89	16.97	17.06	17.14	17.22	17.31	17.39	17.47	17.56	
2	17.64	17.72	17.81	17.89	17.98	18.06	18.15	18.23	18.32	18.40	
3	18.49	18.58	18.66	18.75	18.84	18.92	19.01	19.10	19.18	19.27	9
4	19.36	19.45	19.54	19.62	19.71	19.80	19.89	19.98	20.07	20.16	
4.5	20.25	20.34	20.43	20.52	20.61	20.70	20.79	20.88	20.98	21.07	
6	21.16	21.25	21.34	21.44	21.53	21.62	21.72	21.81	21.90	22.00	
7	22.09	22.18	22.28	22.37	22.47	22.56	22.66	22.75	22.85	22.94	10
8	23.04	23.14	23.23	23.33	23.43	23.52	23.62	23.72	23.81	23.91	
9	24.01	24.11	24.21	24.30	24.40	24.50	24.60	24.70	24.80	24.90	

$\pi^2 = 9.86960$ $(\pi/2)^2 = 2.46740$ $1/\pi^2 = 0.101321$

Explanation of Table of Squares.

This table gives the value of N^2 for values of N from 1 to 10, correct to four figures. (Interpolated values may be in error by 1 in the fourth figure).

To find the square of a number N outside the range from 1 to 10, note that moving the decimal point one place in column N is equivalent to moving it **two places** in the body of the table. For example:

$(3.217)^2 = 10.35$; $(0.03217)^2 = 0.001035$; $(3217)^2 = 10350000$

This table can also be used inversely, to give square roots.

SQUARES *(continued)*

N	0	1	2	3	4	5	6	7	8	9	Avg. diff.
5.0	25.00	25.10	25.20	25.30	25.40	25.50	25.60	25.70	25.81	25.91	10
1	26.01	26.11	26.21	26.32	26.42	26.52	26.63	26.73	26.83	26.94	
2	27.04	27.14	27.25	27.35	27.46	27.56	27.67	27.77	27.88	27.98	
3	28.09	28.20	28.30	28.41	28.52	28.62	28.73	28.84	28.94	29.05	11
4	29.16	29.27	29.38	29.48	29.59	29.70	29.81	29.92	30.03	30.14	
5.5	30.25	30.36	30.47	30.58	30.69	30.80	30.91	31.02	31.14	31.25	
6	31.36	31.47	31.58	31.70	31.81	31.92	32.04	32.15	32.26	32.38	
7	32.49	32.60	32.72	32.83	32.95	33.06	33.18	33.29	33.41	33.52	
8	33.64	33.76	33.87	33.99	34.11	34.22	34.34	34.46	34.57	34.69	12
9	34.81	34.93	35.05	35.16	35.28	35.40	35.52	35.64	35.76	35.88	
6.0	36.00	36.12	36.24	36.36	36.48	36.60	36.72	36.84	36.97	37.09	
1	37.21	37.33	37.45	37.58	37.70	37.82	37.95	38.07	38.19	38.32	
2	38.44	38.56	38.69	38.81	38.94	39.06	39.19	39.31	39.44	39.56	
3	39.69	39.82	39.94	40.07	40.20	40.32	40.45	40.58	40.70	40.83	13
4	40.96	41.09	41.22	41.34	41.47	41.60	41.73	41.86	41.99	42.12	
6.5	42.25	42.38	42.51	42.64	42.77	42.90	43.03	43.16	43.30	43.43	
6	43.56	43.69	43.82	43.96	44.09	44.22	44.36	44.49	44.62	44.76	
7	44.89	45.02	45.16	45.29	45.43	45.56	45.70	45.83	45.97	46.10	
8	46.24	46.38	46.51	46.65	46.79	46.92	47.06	47.20	47.33	47.47	14
9	47.61	47.75	47.89	48.02	48.16	48.30	48.44	48.58	48.72	48.86	
7.0	49.00	49.14	49.28	49.42	49.56	49.70	49.84	49.98	50.13	50.27	
1	50.41	50.55	50.69	50.84	50.98	51.12	51.27	51.41	51.55	51.70	
2	51.84	51.98	52.13	52.27	52.42	52.56	52.71	52.85	53.00	53.14	
3	53.29	53.44	53.58	53.73	53.88	54.02	54.17	54.32	54.46	54.61	15
4	54.76	54.91	55.06	55.20	55.35	55.50	55.65	55.80	55.95	56.10	
7.5	56.25	56.40	56.55	56.70	56.85	57.00	57.15	57.30	57.46	57.61	
6	57.76	57.91	58.06	58.22	58.37	58.52	58.68	58.83	58.98	59.14	
7	59.29	59.44	59.60	59.75	59.91	60.06	60.22	60.37	60.53	60.68	
8	60.84	61.00	61.15	61.31	61.47	61.62	61.78	61.94	62.09	62.25	16
9	62.41	62.57	62.73	62.88	63.04	63.20	63.36	63.52	63.68	63.84	
8.0	64.00	64.16	64.32	64.48	64.64	64.80	64.96	65.12	65.29	65.45	
1	65.61	65.77	65.93	66.10	66.26	66.42	66.59	66.75	66.91	67.08	
2	67.24	67.40	67.57	67.73	67.90	68.06	68.23	68.39	68.56	68.72	
3	68.89	69.06	69.22	69.39	69.56	69.72	69.89	70.06	70.22	70.39	17
4	70.56	70.73	70.90	71.06	71.23	71.40	71.57	71.74	71.91	72.08	
8.5	72.25	72.42	72.59	72.76	72.93	73.10	73.27	73.44	73.62	73.79	
6	73.96	74.13	74.30	74.48	74.65	74.82	75.00	75.17	75.34	75.52	
7	75.69	75.86	76.04	76.21	76.39	76.56	76.74	76.91	77.09	77.26	
8	77.44	77.62	77.79	77.97	78.15	78.32	78.50	78.68	78.85	79.03	18
9	79.21	79.39	79.57	79.74	79.92	80.10	80.28	80.46	80.64	80.82	
9.0	81.00	81.18	81.36	81.54	81.72	81.90	82.08	82.26	82.45	82.63	
1	82.81	82.99	83.17	83.36	83.54	83.72	83.91	84.09	84.27	84.46	
2	84.64	84.82	85.01	85.19	85.38	85.56	85.75	85.93	86.12	86.30	
3	86.49	86.68	86.86	87.05	87.24	87.42	87.61	87.80	87.98	88.17	19
4	88.36	88.55	88.74	88.92	89.11	89.30	89.49	89.68	89.87	90.06	
9.5	90.25	90.44	90.63	90.82	91.01	91.20	91.39	91.58	91.78	91.97	
6	92.16	92.35	92.54	92.74	92.93	93.12	93.32	93.51	93.70	93.90	
7	94.09	94.28	94.48	94.67	94.87	95.06	95.26	95.45	95.65	95.84	
8	96.04	96.24	96.43	96.63	96.83	97.02	97.22	97.42	97.61	97.81	20
9	98.01	98.21	98.41	98.60	98.80	99.00	99.20	99.40	99.60	99.80	
10.0	100.0										

Moving the decimal point ONE place in *N* requires moving it TWO places in body of table.

COMMON LOGARITHMS (*special table*)

Number	0	1	2	3	4	5	6	7	8	9	Avg. dif.
1.00	0.0000	0004	0009	0013	0017	0022	0026	0030	0035	0039	4
1.01	0043	0048	0052	0056	0060	0065	0069	0073	0077	0082	
1.02	0086	0090	0095	0099	0103	0107	0111	0116	0120	0124	
1.03	0128	0133	0137	0141	0145	0149	0154	0158	0162	0166	
1.04	0170	0175	0179	0183	0187	0191	0195	0199	0204	0208	
1.05	0212	0216	0220	0224	0228	0233	0237	0241	0245	0249	
1.06	0253	0257	0261	0265	0269	0273	0278	0282	0286	0290	
1.07	0294	0298	0302	0306	0310	0314	0318	0322	0326	0330	
1.08	0334	0338	0342	0346	0350	0354	0358	0362	0366	0370	
1.09	0374	0378	0382	0386	0390	0394	0398	0402	0406	0410	
1.10	0.0414	0418	0422	0426	0430	0434	0438	0441	0445	0449	
1.11	0453	0457	0461	0465	0469	0473	0477	0481	0484	0488	
1.12	0492	0496	0500	0504	0508	0512	0515	0519	0523	0527	
1.13	0531	0535	0538	0542	0546	0550	0554	0558	0561	0565	
1.14	0569	0573	0577	0580	0584	0588	0592	0596	0599	0603	
1.15	0607	0611	0615	0618	0622	0626	0630	0633	0637	0641	
1.16	0645	0648	0652	0656	0660	0663	0667	0671	0674	0678	
1.17	0682	0686	0689	0693	0697	0700	0704	0708	0711	0715	
1.18	0719	0722	0726	0730	0734	0737	0741	0745	0748	0752	
1.19	0755	0759	0763	0766	0770	0774	0777	0781	0785	0788	
1.20	0.0792	0795	0799	0803	0806	0810	0813	0817	0821	0824	
1.21	0828	0831	0835	0839	0842	0846	0849	0853	0856	0860	
1.22	0864	0867	0871	0874	0878	0881	0885	0888	0892	0896	
1.23	0899	0903	0906	0910	0913	0917	0920	0924	0927	0931	
1.24	0934	0938	0941	0945	0948	0952	0955	0959	0962	0966	
1.25	0969	0973	0976	0980	0983	0986	0990	0993	0997	1000	3
1.26	1004	1007	1011	1014	1017	1021	1024	1028	1031	1035	
1.27	1038	1041	1045	1048	1052	1055	1059	1062	1065	1069	
1.28	1072	1075	1079	1082	1086	1089	1092	1096	1099	1103	
1.29	1106	1109	1113	1116	1119	1123	1126	1129	1133	1136	
1.30	0.1139	1143	1146	1149	1153	1156	1159	1163	1166	1169	
1.31	1173	1176	1179	1183	1186	1189	1193	1196	1199	1202	
1.32	1206	1209	1212	1216	1219	1222	1225	1229	1232	1235	
1.33	1239	1242	1245	1248	1252	1255	1258	1261	1265	1268	
1.34	1271	1274	1278	1281	1284	1287	1290	1294	1297	1300	
1.35	1303	1307	1310	1313	1316	1319	1323	1326	1329	1332	
1.36	1335	1339	1342	1345	1348	1351	1355	1358	1361	1364	
1.37	1367	1370	1374	1377	1380	1383	1386	1389	1392	1396	
1.38	1399	1402	1405	1408	1411	1414	1418	1421	1424	1427	
1.39	1430	1433	1436	1440	1443	1446	1449	1452	1455	1458	
1.40	0.1461	1464	1467	1471	1474	1477	1480	1483	1486	1489	
1.41	1492	1495	1498	1501	1504	1508	1511	1514	1517	1520	
1.42	1523	1526	1529	1532	1535	1538	1541	1544	1547	1550	
1.43	1553	1556	1559	1562	1565	1569	1572	1575	1578	1581	
1.44	1584	1587	1590	1593	1596	1599	1602	1605	1608	1611	
1.45	1614	1617	1620	1623	1626	1629	1632	1635	1638	1641	
1.46	1644	1647	1649	1652	1655	1658	1661	1664	1667	1670	
1.47	1673	1676	1679	1682	1685	1688	1691	1694	1697	1700	
1.48	1703	1706	1708	1711	1714	1717	1720	1723	1726	1729	
1.49	1732	1735	1738	1741	1744	1746	1749	1752	1755	1758	

Moving the decimal point n places to the right [or left] in the number requires adding $+n$ [or $-n$] in the body of the table.

COMMON LOGARITHMS (*special table, continued*)

Number	0	1	2	3	4	5	6	7	8	9	Avg. diff.
1.50	0.1761	1764	1767	1770	1772	1775	1778	1781	1784	1787	3
1.51	1790	1793	1796	1798	1801	1804	1807	1810	1813	1816	
1.52	1818	1821	1824	1827	1830	1833	1836	1838	1841	1844	
1.53	1847	1850	1853	1855	1858	1861	1864	1867	1870	1872	
1.54	1875	1878	1881	1884	1886	1889	1892	1895	1898	1901	
1.55	1903	1906	1909	1912	1915	1917	1920	1923	1926	1928	
1.56	1931	1934	1937	1940	1942	1945	1948	1951	1953	1956	
1.57	1959	1962	1965	1967	1970	1973	1976	1978	1981	1984	
1.58	1987	1989	1992	1995	1998	2000	2003	2006	2009	2011	
1.59	2014	2017	2019	2022	2025	2028	2030	2033	2036	2038	
1.60	0.2041	2044	2047	2049	2052	2055	2057	2060	2063	2066	
1.61	2068	2071	2074	2076	2079	2082	2084	2087	2090	2092	
1.62	2095	2098	2101	2103	2106	2109	2111	2114	2117	2119	
1.63	2122	2125	2127	2130	2133	2135	2138	2148	2143	2146	
1.64	2148	2151	2154	2156	2159	2162	2164	2167	2170	2172	
1.65	2175	2177	2180	2183	2185	2188	2191	2193	2196	2198	
1.66	2201	2204	2206	2209	2212	2214	2217	2219	2222	2225	
1.67	2227	2230	2232	2235	2238	2240	2243	2245	2248	2251	
1.68	2253	2256	2258	2261	2263	2266	2269	2271	2274	2276	
1.69	2279	2281	2284	2287	2289	2292	2294	2297	2299	2302	
1.70	0.2304	2307	2310	2312	2315	2317	2320	2322	2325	2327	
1.71	2330	2333	2335	2338	2340	2343	2345	2348	2350	2353	
1.72	2355	2358	2360	2363	2365	2368	2370	2373	2375	2378	
1.73	2380	2383	2385	2388	2390	2393	2395	2398	2400	2403	
1.74	2405	2408	2410	2413	2415	2418	2420	2423	2425	2428	2
1.75	2430	2433	2435	2438	2440	2443	2445	2448	2450	2453	
1.76	2455	2458	2460	2463	2465	2467	2470	2472	2475	2477	
1.77	2480	2482	2485	2487	2490	2492	2494	2497	2499	2502	
1.78	2504	2507	2509	2512	2514	2516	2519	2521	2524	2526	
1.79	2529	2531	2533	2536	2538	2541	2543	2545	2548	2550	
1.80	0.2553	2555	2558	2560	2562	2565	2567	2570	2572	2574	
1.81	2577	2579	2582	2584	2586	2589	2591	2594	2596	2598	
1.82	2601	2603	2605	2608	2610	2613	2615	2617	2620	2622	
1.83	2625	2627	2629	2632	2634	2636	2639	2641	2643	2646	
1.84	2648	2651	2653	2655	2658	2660	2662	2665	2667	2669	
1.85	2672	2674	2676	2679	2681	2683	2686	2688	2690	2693	
1.86	2695	2697	2700	2702	2704	2707	2709	2711	2714	2716	
1.87	2718	2721	2723	2725	2728	2730	2732	2735	2737	2739	
1.88	2742	2744	2746	2749	2751	2753	2755	2758	2760	2762	
1.89	2765	2767	2769	2772	2774	2776	2778	2781	2783	2785	
1.90	0.2788	2790	2792	2794	2797	2799	2801	2804	2806	2808	
1.91	2810	2813	2815	2817	2819	2822	2824	2826	2828	2831	
1.92	2833	2835	2838	2840	2842	2844	2847	2849	2851	2853	
1.93	2856	2858	2860	2862	2865	2867	2869	2871	2874	2876	
1.94	2878	2880	2882	2885	2887	2889	2891	2894	2896	2898	
1.95	2900	2903	2905	2907	2909	2911	2914	2916	2918	2920	
1.96	2923	2925	2927	2929	2931	2934	2936	2938	2940	2942	
1.97	2945	2947	2949	2951	2953	2956	2958	2960	2962	2964	
1.98	2967	2969	2971	2973	2975	2978	2980	2982	2984	2986	
1.99	2989	2991	2993	2995	2997	2999	3002	3004	3006	3008	

COMMON LOGARITHMS

Number	0	1	2	3	4	5	6	7	8	9	Avg. diff.
1.0	0.0000	0043	0086	0128	0170	0212	0253	0294	0334	0374	
1.1	0414	0453	0492	0531	0569	0607	0645	0682	0719	0755	
1.2	0792	0828	0864	0899	0934	0969	1004	1038	1072	1106	
1.3	1139	1173	1206	1239	1271	1303	1335	1367	1399	1430	
1.4	1461	1492	1523	1553	1584	1614	1644	1673	1703	1732	
1.5	1761	1790	1818	1847	1875	1903	1931	1959	1987	2014	
1.6	2041	2068	2095	2122	2148	2175	2201	2227	2253	2279	
1.7	2304	2330	2355	2380	2405	2430	2455	2480	2504	2529	
1.8	2553	2577	2601	2625	2648	2672	2695	2718	2742	2765	
1.9	2788	2810	2833	2856	2878	2900	2923	2945	2967	2989	
2.0	0.3010	3032	3054	3075	3096	3118	3139	3160	3181	3201	21
2.1	3222	3243	3263	3284	3304	3324	3345	3365	3385	3404	20
2.2	3424	3444	3464	3483	3502	3522	3541	3560	3579	3598	19
2.3	3617	3636	3655	3674	3692	3711	3729	3747	3766	3784	18
2.4	3802	3820	3838	3856	3874	3892	3309	3927	3945	3962	17
2.5	3979	3997	4014	4031	4048	4065	4082	4099	4116	4133	17
2.6	4150	4166	4183	4200	4216	4232	4249	4265	4281	4298	16
2.7	4314	4330	4346	4362	4378	4393	4409	4425	4440	4456	16
2.8	4472	4487	4502	4518	4533	4548	4564	4579	4594	4609	15
2.9	4624	4639	4654	4669	4683	4698	4713	4728	4742	4757	15
3.0	0.4771	4786	4800	4814	4829	4843	4857	4871	4886	4900	14
3.1	4914	4928	4942	4955	4969	4983	4997	5011	5024	5038	14
3.2	5051	5065	5079	5092	5105	5119	5132	5145	5159	5172	13
3.3	5185	5198	5211	5224	5237	5250	5263	5276	5289	5302	13
3.4	5315	5328	5340	5353	5366	5378	5391	5403	5416	5428	13
3.5	5441	5453	5465	5478	5490	5502	5514	5527	5539	5551	12
3.6	5563	5575	5587	5599	5611	5623	5635	5647	5658	5670	12
3.7	5682	5694	5705	5717	5729	5740	5752	5763	5775	5786	12
3.8	5798	5809	5821	5832	5843	5855	5866	5877	5888	5899	11
3.9	5911	5922	5933	5944	5955	5966	5977	5988	5999	6010	11
4.0	0.6021	6031	6042	6053	6064	6075	6085	6096	6107	6117	11
4.1	6128	6138	6149	6160	6170	6180	6191	6201	6212	6222	10
4.2	6232	6243	6253	6263	6274	6284	6294	6304	6314	6325	10
4.3	6335	6345	6355	6365	6375	6385	6395	6405	6415	6425	10
4.4	6435	6444	6454	6464	6474	6484	6493	6503	6513	6522	10
4.5	6532	6542	6551	6561	6571	6580	6590	6599	6609	6618	10
4.6	6628	6637	6646	6656	6665	6675	6684	6693	6702	6712	10
4.7	6721	6730	6739	6749	6758	6767	6776	6785	6794	6803	9
4.8	6812	6821	6830	6839	6848	6857	6866	6875	6884	6893	9
4.9	6902	6911	6920	6928	6937	6946	6955	6964	6972	6981	9

$\log \pi = 0.4971$ $\log \pi/2 = 0.1961$ $\log \pi^2 = 0.9943$ $\log \sqrt{\pi} = 0.2486$
$\log e = 0.4343$ $\log (0.4343) = 0.6378 - 1$

These two pages give the common logarithms of numbers between 1 and 10, *correct to four places.* Moving the decimal point n places to the right [*or* left] in the number is equivalent to adding n [*or* $-n$] to the logarithm. Thus, $\log 0.017453 = 0.2419 - 2$, which may also be written $\bar{2}.2419$ or $8.2419 - 10$.

$\log (ab) = \log a + \log b$ $\log (a^N) = N \log a$

$\log \left(\frac{a}{b}\right) = \log a - \log b$ $\log (\sqrt[N]{a}) = \frac{1}{N} \log a$

COMMON LOGARITHMS (*continued*)

Number	0	1	2	3	4	5	6	7	8	9	Avg. dif.
5.0	0.6990	6998	7007	7016	7024	7033	7042	7050	7059	7067	9
5.1	7076	7084	7093	7101	7110	7118	7126	7135	7143	7152	8
5.2	7160	7168	7177	7185	7193	7202	7210	7218	7226	7235	8
5.3	7243	7251	7259	7267	7275	7284	7292	7300	7308	7316	8
5.4	7324	7332	7340	7348	7356	7364	7372	7380	7388	7396	8
5.5	7404	7412	7419	7427	7435	7443	7451	7459	7466	7474	8
5.6	7482	7490	7497	7505	7513	7520	7528	7536	7543	7551	8
5.7	7559	7566	7574	7582	7589	7597	7604	7612	7619	7627	8
5.8	7634	7642	7649	7657	7664	7672	7679	7686	7694	7701	7
5.9	7709	7716	7723	7731	7738	7745	7752	7760	7767	7774	7
6.0	0.7782	7789	7796	7803	7810	7818	7825	7832	7839	7846	7
6.1	7853	7860	7868	7875	7882	7889	7896	7903	7910	7917	7
6.2	7924	7931	7938	7945	7952	7959	7966	7973	7980	7987	7
6.3	7993	8000	8007	8014	8021	8028	8035	8041	8048	8055	7
6.4	8062	8069	8075	8082	8089	8096	8102	8109	8116	8122	7
6.5	8129	8136	8142	8149	8156	8162	8169	8176	8182	8189	7
6.6	8195	8202	8209	8215	8222	8228	8235	8241	8248	8254	7
6.7	8261	8267	8274	8280	8287	8293	8299	8306	8312	8319	6
6.8	8325	8331	8338	8344	8351	8357	8363	8370	8376	8382	6
6.9	8388	8395	8401	8407	8414	8420	8426	8432	8439	8445	6
7.0	0.8451	8457	8463	8470	8476	8482	8488	8494	8500	8506	6
7.1	8513	8519	8525	8531	8537	8543	8549	8555	8561	8567	6
7.2	8573	8579	8585	8591	8597	8603	8609	8615	8621	8627	6
7.3	8633	8639	8645	8651	8657	8663	8669	8675	8681	8686	6
7.4	8692	8698	8704	8710	8716	8722	8727	8733	8739	8745	6
7.5	8751	8756	8762	8768	8774	8779	8785	8791	8797	8802	6
7.6	8808	8814	8820	8825	8831	8837	8842	8848	8854	8859	6
7.7	8865	8871	8876	8882	8887	8893	8899	8904	8910	8915	6
7.8	8921	8927	8932	8938	8943	8949	8954	8960	8965	8971	6
7.9	8976	8982	8987	8993	8998	9004	9009	9015	9020	9025	5
8.0	0.9031	9036	9042	9047	9053	9058	9063	9069	9074	9079	5
8.1	9085	9090	9096	9101	9106	9112	9117	9122	9128	9133	5
8.2	9138	9143	9149	9154	9159	9165	9170	9175	9180	9186	5
8.3	9191	9196	9201	9206	9212	9217	9222	9227	9232	9238	5
8.4	9243	9248	9253	9258	9263	9269	9274	9279	9284	9289	5
8.5	9294	9299	9304	9309	9315	9320	9325	9330	9335	9340	5
8.6	9345	9350	9355	9360	9365	9370	9375	9380	9385	9390	5
8.7	9395	9400	9405	9410	9415	9420	9425	9430	9435	9440	5
8.8	9445	9450	9455	9460	9465	9469	9474	9479	9484	9489	5
8.9	9494	9499	9504	9509	9513	9518	9523	9528	9533	9538	5
9.0	0.9542	9547	9552	9557	9562	9566	9571	9576	9581	9586	5
9.1	9590	9595	9600	9605	9609	9614	9619	9624	9628	9633	5
9.2	9638	9643	9647	9652	9657	9661	9666	9671	9675	9680	5
9.3	9685	9689	9694	9699	9703	9708	9713	9717	9722	9727	5
9.4	9731	9736	9741	9745	9750	9754	9759	9763	9768	9773	5
9.5	9777	9782	9786	9791	9795	9800	9805	9809	9814	9818	5
9.6	9823	9827	9832	9836	9841	9845	9850	9854	9859	9863	4
9.7	9868	9872	9877	9881	9886	9890	9894	9899	9903	9908	4
9.8	9912	9917	9921	9926	9930	9934	9939	9943	9948	9952	4
9.9	9956	9961	9965	9969	9974	9978	9983	9987	9991	9996	4

INDEX

Numbers refer to pages.

A

Accuracy, degree of, 8
  false, 8
Area, determined by definite integral, 341
  of frequency polygon, 74
  of rectangular histogram, 73
  under curve, 76
Array, 4
Average, 96 (*see also* Mean),
  for comparison, 3
  general definition of, 96
  rate of growth, 124
  representative type form, 96
Axis, 56
  of abscissas, 56
  of dependent variable, 56
  of independent variable, 56
  of ordinates, 56
  of parabola, 65

B

Base of a system of logarithms, 295
Binomial distribution, non-symmetric, 35
    symmetric, 33
      freshman heights compared with, 33, 86
    histograms, 77
  expansion, formula for, 310
    if *n* is not a positive integer, 313
    rules for forming terms of, 312
Boundaries, class, 25
  distinction of, from class limits, 27

C

Characteristic of logarithms, 299
  rule for finding, 299
Charlier check, 233
Chart (*see also* Diagram),
  geometric, 254
  ratio, 256
    shows per cent change, 261
Circle, equation of, 63
Class, 5
  boundaries, 25
  frequency, 5
  interval, 5
    as unit of measure, 232
    effect of widening, 27, 84
  limits, 25
    distribution from class boundaries, 27
Closest fit, 186
  frequency curve of, 200
  logarithmic least squares line of, 264
  parabola of, 196
    normal equations for, 197
  straight line of, 187
    graphical method for, 187
    least squares, 190
    method of averaging for, 188
    normal equations for least squares, 192
Cologarithm, 305
Combinations, definition of, 307
  number of, of *n* things *r* at a time, 309
  simple relationship between, 309
Comparison, as purpose of statistics, 3
Constant of integration, 338
  cancels in definite integral, 342
Coordinates, 56
  axes of, 56
  three, 333
Correlation, Chap. IX
  by ranks, 243

Correlation, definition of, 212
formula for Pearson's coefficient of, 229
origin at zero, 231
short-cut, 229
idea of, developed through probability, 214–222
measure of, 227, 228
ratio, 236
skew, 227, 236
table, construction of, 231
form of, 234
when r equals zero, 240
Curve fitting, Chap. VIII
by normal probability curve, 200, 201, 202
by parabola, 196
by straight line, 187
certain type forms for, 186

D

Deck, 255
Derivative, 325
equals slope of tangent to a curve, 325
speed, 332
of x^n, 325
partial, 334
second, 330
equals acceleration, 332
Deviation, 5, 7
average, 134
quartile, 159
standard, 137
computation of, from frequency table, 142
equals $\sqrt{npq}$, 147
measure of consistency, 139
mechanical interpretation of, 139
other names for, 139
satisfies general definition of average, 147
short-cut computation of, 144
Deviations, sum of, from any arbitrary value, 98
from arithmetic mean, 97
Diagram, bar, 42, 44
Diagram, line, 51
one hundred per cent bar, 44
peak-top, 51
pie, 45
percentage, 47
use of, 47
scatter, 39, 213
Dispersion, coefficient of, 135, 141, 159
comparison of measures of, 161
for comparison, 4
idea of, 131
measure of, average deviation as, 134
range as, 132
standard deviation as, 137
quartile measure of, 158
Division points, 150
formulas for, 155
graphical determination of, 161
in general, 153
median, 111, 151
percentiles, 153
quartiles, 150, 151, 152
significance of, 153
Double entry, table of, 35, 36

E

Equation, of circle, 63
normal probability curve, 177, 185, 200
parabola, 64
of closest fit, 196, 197
straight-line, 57, 59, 60, 62
of closest fit, 187, 188, 190, 222
regression lines, 222, 224, 225

F

Forecasting, 260, 266
Frequency, 25
classes, 160
curve, 74
area of, 76
drawing of, 75
fitting of, 76, 84, 200, 204
symmetric binomial, 77
distributions, comparison of, 29
due to pure chance, 29
polygon, 73
area of, 74

Frequency, tables, 39
  theoretical, in tossing coins, 31
Function, 50, 324
  derivative of, 325
    gives slope of tangent, 325
  graph of, 51
  increasing and decreasing, 329
    linear, 61
  maximum value of, 328
    test for, 330
  minimum value of, 329
    test for, 330
  of a function, derivative of, 332
    rule for, 333
  of two variables, 333
    application to least squares adjustment, 337
    graph of, 334
    maximum or minimum values of, 335
  plotting of, 53
  second degree, 65

G

Geometric mean, 122
  as average rate of change, 265
  connection, with arithmetic mean, 123
    with harmonic mean, 127
  slope of logarithmic straight line, 266
Gompers curve, 267
Graduation, 198
  of frequency distribution, 200
  of freshman heights to normal probability curve, 204
Graph, area of, 73, 74, 76 (*see also* Diagram).
  frequency, 71
  general rules for making, 89
  logarithmic, Chap. X
    reading of scale numbers on, 255
    unlimited, 255
  object of, 42

H

Heads and tails, number of in 500 throws of seven dimes, 16
Histogram, 71
Histogram, from pure chance, 76
  rectangular, 72
    area of, 73, 74
  symmetric binomial, 77
    comparison of, with freshman heights, 86

I

Index numbers, Chap. XI
  as type forms, 293
  average price relative used as, 277
  averages used in computation of, 280
  bias in formulas for, 288
  crossing of formulas for, 288
  definition of, 276
  Fisher's "Ideal," 289
  forms of computation of, 290
  notation used in formulas, 276
  simple, 275
  systems of weighing used for, 281
  tests, of good formulas for, 282
    commodity reversal, 282
    factor reversal, 283
    time reversal, 282
  warning as to use of, 292
Integration, between limits, 341
  constant of, 338
  definition of, 338
  geometric meaning of, 339
Interpolation, by A. M. and G. M., 124
  graphical, 51, 52
  in table of logarithms, 301

K

Kurtosis, 206

L

Least squares adjustment, 190
  formulas, derivation of. 337
    for parabola of, 197
    for straight line of, 192
  logarithmic line of, 200, 264
Limit of a variable, 321
  illustration of, reached by the variable, 321
    that cannot be reached by the variable, 321

Limits, class, 25
of definite integral, *341, 342*
Logarithms, base of system of, 295
other than 10, 303
ten, 297
characteristic of, 298
rule for, 299
definition of, 295
interpolation in table of, 301
laws of, 295
mantissa of, 297
independent of decimal point, 300
modulus of system of, 304
natural or Napierian, 304
Lorenz curve, 162
interpretation of, 164

M

Mantissa, 297
independent of decimal point, 300
Maps, statistical, 89
Maximum value of a function, 328
test for, 330
Mean, arithmetic, 96
determination by frequency curve, 119
determination by frequency table, 99
short-cut, 100
important property of, 97
not always representative, 102
satisfies general definition of average, 97
short-cut for computing, 98
sum of deviations from equals zero, 97
weighted, 121
geometric, 122
connection with A. M., 123
graphical determination, 265
rate of growth by, 124
harmonic, 126
connection between A. M. and G. M., 127
satisfies general definition of average, 128
Median, 111
definition of, more precise, 112, 151
differences from, sum of, a minimum, 117
formula for, by interpolation, 114, 155
frequency curve, determination by, 119
ogive, determination by, 116
satisfies general definition of average, 118
Minimum value of a function, 329
test for, 330
Mode, 103
effect on, by shifting class limits, 110
formula for, 103
frequency graph, determined by, 103
grouping, determination by, 105, 107
ogive, determination by, 109
Moving average, 207
graph of, or line of trend, 209

N

Normal equations, for parabola of closest fit, 197
for straight line of closest fit, 192
other type forms of, 199
probability curve, 82, 177, 185, 200, 201

O

Ogive, 87
comparison of, with histogram, 88
division points determined by, 161
median determined by, 116
mode determined by, 109
Organic growth, law of, 257

P

Parabola, equation of, 64
of closest fit, 196
normal equations for, 197
use of, 65

Pareto's law, 268
Pascal's triangle, 312
Permutations, definitions of, 307
  value of, of n things r at a time, 308
Pictogram, 48
  use with caution, 50
Price relative, 277
  average, used as index number, 277
Probability, compound, 315
  definition of, 315
  laws of, 30
    proof of, 315–320
  normal curve of, 82, 177, 185, 200, 201
    area under, 178
  partial, 317
  total, 318
  used to develop idea of correlation, 214–222
Probable error, definition of, 177
  equals 0.674 of standard deviation, 179, 180, 203
  examples of use of, 182
  odds against a random selection being beyond three times, 181
  of various statistical constants, 181
  value of, depends on N, 183

Q

Quartile, 150
  coefficient of dispersion, 159
  first, 151
  measure of dispersion, 159
  third, 152

R

Ranks, correlation by, 243
    derivation of formula for, 244
  definition of, 243
Regression coefficients, 225, 226
    product of, a measure of correlation, 227
  equations, 222 to 225
    referred to axes through (0,0), 240
  linear, 227
    departure from, 239
Regression lines, 222, 225
    and correlation, 227
    coincidence of, 227
    meaning of, 225
    on coordinate axes, 227
    plotting of, 242

S

Sampling, 4, 74
Saturation point, 258, 261
Scale, choosing of, 89, 90
  logarithmic, 68, 252
  marks to be on heavy lines, 90
  of squares, 66
  uniform and non-uniform, 66
Scatter diagram, 39, 213
Skewness, 169
  coefficient of, 171, 172
  effect on mode, median, and arithmetic mean, 169
  extreme forms of, 175
  limits on, in fitting normal curve, 206
  measure of, 170, 172
  significant degree of, 175, 183
Slope of logarithmic straight line of closest fit, 266
Smoothing, 52, 75, 76
Standard deviation, 137
  chances of a random choice falling beyond certain multiples of, 179, 203
  computation of, direct method, 142
    short-cut method, 144
  equals $\sqrt{npq}$, 147
  is an average, 147
  measure of consistency, 139
  mechanical interpretation of, 139
  other names for, 139
Statistical regularity, law of, 4
Statistics, definitions of, dictionary, 1
  Yule's, 2
Straight line, equation of, intercept form of, 60
    not through the origin, 59
    through the origin, 57
      two points, 26

Straight line of closest fit, 187
graphical method of finding, 187
least squares, 190
formulas derived for, 192
principle involved, 190
simplified formulas for, 193
method of averaging, 188
normal equations for, 192
slope of, 58

T

Tabulation, Chap. *II*
Tangent to a curve, definition of, 323
slope of, 324
equals derivative of function, 325
Trend, line of, by moving average, 209
logarithmic least squares line of, 264

V

Variable, continuous, 3
definition, 3
dependent, 50
discrete, 3
historical, 3
independent, 50
not reaching limit, 321
reaching the limit, 321
Variates, definition of, 3
Variation, time, place, quality, 23, 24

AGRICULTURAL ... RCH INSTITUTE LIBRARY
No. 1464
9-NOV 1926
COIMBATORE